National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

UNIVERSITY OF ALBERTA

APPLICATION OF GENERALIZABILITY THEORY TO THE ANALYSIS OF DESIGNS INVOLVING RANDOM NESTED DEPENDENT VARIABLES

BY

RANDY ELDER

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of MASTER OF EDUCATION.

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

Edmonton, Alberta

Fall, 1991

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-70252-4

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: RANDY W. ELDER

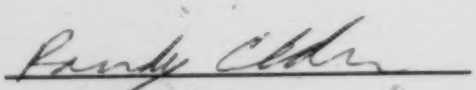
TITLE OF THESIS: Application of generalizability theory to the analysis of designs involving random nested dependent variables.

DEGREE: Master of Education

YEAR THIS DEGREE GRANTED: 1991

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



6714-105A Street
Edmonton, Alberta
T6H 2R2

September 30, 1991

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis titled APPLICATION OF GENERALIZABILITY THEORY TO THE ANALYSIS OF DESIGNS INVOLVING RANDOM NESTED DEPENDENT VARIABLES submitted by RANDY ELDER in partial fulfillment of the requirements for the degree of MASTER OF EDUCATION.

William Todd Rogers

Steve Hunka

C. Don Heth

September 30, 1991

ABSTRACT

The value of generalizability theory as a perspective from which to conduct research is discussed as a means of assessing the dependent measure and encouraging logical consistency within the study. Adopting this approach illuminates the logical necessity of the treatment of dependent variables as random factors to justify the desired inferences beyond the particular conditions used in many educational and psychological studies. Treatment of dependent variables as random results in difficulties in data analysis due to the resulting lack of an appropriate error term for F-tests in many designs, however. Appropriate methods of data analysis for this situation as discussed in Hopkins (1984) are illustrated.

The use of the approach advocated by Hopkins is extended on a typical data set found in education, the results of the June, 1989 British Columbia Algebra 12 school leaving examinations. In this data set, items are nested within content domains and within levels of cognitive complexity, resulting in an unbalanced design due to different numbers of items being nested within each blocking variable.

Problems in the estimation of variance components resulting from such unbalancing are discussed and computational methods currently available to provide variance estimates from unbalanced data are assessed. Henderson's Method III (Type I) was the method of choice for this study and it was utilized to analyze the Algebra 12 data for four designs of varying complexity. Due to the large

design matrices resulting from the inclusion of persons as a factor in generalizability theory, the computing resources required for an analysis became a primary consideration in the choice of method. Although each available method of analysis has advantages and disadvantages, it is clear that further research into accurate and efficient methods of variance component estimation for unbalanced designs would be valuable.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Dr. W.T. Rogers for his patience, support, and guidance throughout the course of this study and throughout my time at CRAME.

I would also like to thank my other commitee members, Steve Hunka and Don Heth for their cogent criticisms and for expressing interest in my work.

Thanks to my family and friends; especially those at CRAME for all the stimulating discussions and a sympathetic ear when it was needed.

Finally, thanks to Terry Taerum for assistance in data analysis above and beyond the call of duty.

TABLE OF CONTENTS

CHAPTER I. INTRODUCTION AND OVERVIEW 1
The Nature of the Dependent Variable in Research 2
Statistical Analysis of Designs with Random Dependent Variables 3
Completely Nested Designs 4
Post Hoc Statistical Manipulations 4
An Extension of Hopkins' Design 5
CHAPTER II. LITERATURE REVIEW 8
An Overview of Generalizability Theory 8
The G-Study and Universes of Admissible Observations 11
The Underlying Model for Generalizability Theory 14
An ANOVA Based Method of Estimating Variance Components 16
Application of Variance Components to the Design of Measurement Procedures 17
The D-Study and Universes of Generalization 18
The Structure of D-Study Designs 19
Estimation of Variance Components in D-Studies 20
Fixed Versus Random Facets 21
Universe Scores 21
Universe Score Variance 22

Absolute and Relative Measurement....22
Absolute and Relative Error Variance....24
Generalizability and Dependability Coefficients....24
Issues in the Estimation of Variance Components in Generalizability Theory....26
Sampling Variability and Negative Variance Estimates....27
Unbalanced Designs....29
Choosing a Method of Estimating Variance for Unbalanced Designs....30
Properties and Appropriateness of the Method....30
ANOVA-like Procedures....31
Henderson's Method I....32
Henderson's Method II....32
Henderson's Method III....33
General Approaches....33
Maximum likelihood estimation....34
Minimum norm quadratic unbiased estimation....36
Accuracy of Estimates....37
Computing Resources Required....39
The Fixed-Effect Fallacy....40
Methodological Solutions to the Fixed-effect Fallacy....44
Experimental Design....44

Statistical Manipulations....48
Model simplification (pooling)....48
Quasi F-ratios....50
Synopsis....51
Generalizability Theory....51
Unbalanced Designs....52
The Fixed-Effect Fallacy....53
CHAPTER III. METHODOLOGY....54
Universe of Admissible Observations....54
Universe of Generalization....55
Data Analysis....59
Supplementary analyses....60
Model Simplification and Quasi F-Ratios....60
RESULTS....62
Organization of the Chapter....62
Expected Mean Squares....63
Estimates of Variance Components....69
Assessment of Possible Bias....71
Stability of Variance Estimates....75
Inferential Statistics....77
Model Simplification....77
Quasi F-ratios....81
Generalizability Coefficients....82

CHAPTER V. DISCUSSION 85
Significance Tests for Differences Among Blocking Factors on the Dependent Variable 87
Significance Tests for Differences Among Conditions of the Independent Variable(s) 88
Complexity of Experimental Designs 89
Limitations of the Study 91
New Algorithms for Variance Component Estimation 92
Implications for Future Research 94
Extension to More Complex Designs 94
Implications for Practice 95
REFERENCES 97
APPENDIX A 104
APPENDIX B 107
APPENDIX C 108

LIST OF TABLES

Table 1 Expected Values of Mean Squares Based on the Cornfield & Tukey (1956) Algorithms 16

Table 2 Estimated Variance Components for a p x r x t G-Study Design 17

Table 3 Average C.P.U. Time and Memory Requirements for Estimation of Variance Components for Unbalanced Designs Using Various Methods 41

Table 4 ANOVA Table for the P:T x I Design (I Fixed) 45

Table 5 ANOVA Table for the P:T x I Design (I Random) 46

Table 6 ANOVA Table for the I:P:T Design 46

Table 7 Nesting Structure of the Item Sample From the June, 1989 British Columbia Algebra 12 Examination 58

Table 8 Mean Square Model for the P:G x I Design 64

Table 9 Mean Square Model for the P:G x I:C Design 66

Table 10 Mean Square Model for the P:G x I:L Design 67

Table 11 Mean Square Model for the P:G x I:CL Design 68

Table 12 Variance Estimates Resulting From the Analysis of Four Designs Using Henderson's Method III 72

Table 13 Comparison of Variance Estimates Derived From Method III and ML Analyses for the P:G x I Design 74

Table 14 Comparison of Variance Estimates Derived From Method III and ML Analyses for the P:G x I:C Design on a Reduced Data Set 75

Table 15 Approximate 90% Confidence Intervals for the Variance Estimates Resulting From the Application of Four Models....76
Table 16 ANOVA Table for the Analysis of the P:G x I Design....78
Table 17 ANOVA Table for the Analysis of the P:G x I:C Design....78
Table 18 ANOVA Table for the Analysis of the P:G x I:L Design....79
Table 19 ANOVA Table for the Analysis of the P:G x I:CL Design....78
Table 20 F-ratios for the Gender Effect in Four Experimental Designs....81
Table 21 Quasi F-ratios for the Gender Effect in Four Experimental Designs....82
Table 22 Estimates of Generalizability Coefficients for Various Designs and Methods....84
Table 23 Expected Mean Squares for the Balanced P:G x I Design....104
Table 24 Expected Mean Squares for the Balanced P:G x I:C Design....104
Table 25 Expected Mean Squares for the Balanced P:G x I:L Design....105
Table 26 Expected Mean Squares for the Balanced P:G x I:CL Design....106

CHAPTER I

INTRODUCTION AND OVERVIEW

In classical test theory, educational and psychological measurements are considered to have two components:

1) a 'true score' - the mean of the distribution of a person's scores over an infinite number of hypothetical replications of a measurement, and

2) an error score - the deviation between a person's observed score and true score. This error score is viewed as being made up of random 'noise' from various undifferentiated sources. Generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) was developed as a method of examining this amorphous error more closely in an attempt to clarify the contributions of various specific factors to this error.

Generalizability theory partitions scores into components through the application of a multi-factor ANOVA to the data collected with a measurement instrument. In the simplest case, the variance components associated with each score effect can be estimated by substituting the observed mean squares from the ANOVA into the formulas for the expected mean squares (EMS). By convention, all of the factors in the ANOVA but the objects of measurement are referred to as facets in generalizability theory. However, the symmetry of the ANOVA design must be kept in mind; this implies that any factor in the ANOVA can become the object of measurement (Cardinet, Tourneur, & Allal, 1976). In the

calculation of a generalizability coefficient, which is similar in many respects to the reliability coefficient in classical test theory, the researcher decides which variance components should best be considered to contribute to "true score" variance and which should be considered "error" variance. In generalizability theory, the term true score is often replaced by the term *universe score* to reflect its dependence on the researcher's conception of the universe of generalization.

Although generalizability theory is often strongly associated with the concepts of reliability and validity in measurement, it is actually a powerful technique that can be applied to situations in which random or mixed effects ANOVAs would be performed. The utility of generalizability theory is in the clarification of what is going on beneath the surface of the ANOVA.

The Nature of the Dependent Variable in Research

Using the framework of generalizability theory, Hopkins (1984) demonstrated the incongruity between the design and the statistical analyses in many educational and psychological studies. He pointed out that, in many studies, the total score on an educational or psychological measure is taken as a measure of the dependent variable. In most cases, however, the total score is the sum of scores on individual items. Hopkins argued that if these items are considered to be a random selection from a universe of possible items measuring the same construct, then it follows that "items" should be considered a random factor in a statistical analysis. If items are treated as fixed, then one can only

legitimately generalize the results of a study to that specific set of items, which is clearly not what is intended by most researchers in education and psychology.

Empirical evidence suggests that this failure to consider items as a random facet may be of more than merely theoretical interest. When the significant results of a major series of studies in semantic memory were reexamined with the dependent variable being treated as a random factor, as was appropriate to the generalizations made from the studies, many of these results became suspect (Clark, 1973). However, the incorporation of randomly sampled dependent variables introduces many complexities into studies, particularly those involving crossed factors. It may be for this reason that items have not typically been explicitly treated as random factors in experimental studies.

Statistical Analysis of Designs with Random Dependent Variables

As just indicated, many difficulties are introduced into a statistical analysis when the dependent variable in the analysis is treated as a random sample from the universe of possible measures of that variable. These difficulties arise because when the dependent variable is considered random, there is no longer an appropriate error term for many effects of interest in most experimental designs in which two or more factors are crossed. To address this issue, Hopkins (1984) suggested that there are two basic options available to researchers who wish to legitimately generalize their results beyond the particular items used in their

measurement instrument, the use of nested designs and the use of post hoc statistical manipulations.

Completely Nested Designs

The use of a completely nested design confounds all of the variables in the study and results in an appropriate error term for the main effect when the dependent variable is considered random. Thus, a researcher can simply apply an F-test to the data and arrive at an appropriate result. There is a general consensus that designing studies such that appropriate F-statistics can be obtained is preferable to the post hoc statistical manipulations that will be discussed in the next section (Clark, 1973; Coleman, 1964; Hopkins, 1984; Malgady, Amato, & Huch, 1979). By doing so, however, it is not possible to examine particular interactions of interest because, in contrast to a crossed design, they cannot be isolated.

Post Hoc Statistical Manipulations

In many situations in which it is impossible or undesirable to utilize a completely nested design in a study, the best alternative approach to the data analysis will be to employ post hoc statistical manipulations. Although there are situations in which it is possible to design a study such that an appropriate F-statistic is produced, execution of the design often involves so much effort on the researcher's part that it is not practical or feasible (Malgady et al., 1979). Consequently, the use of post hoc manipulations is advisable in these situations as well.

The two approaches that can be taken to the analysis of a design for which an appropriate error term is not available to test the effect of interest are the use of model simplification and the use of quasi F-ratios. In the model simplification approach, variance components that are likely to have a zero value based on sample data are eliminated from the expected mean squares (EMS) equation for the effect of interest to form simplified EMSs for which an appropriate error term may become available. If model simplification is not justifiable, then quasi F-ratios can be constructed to provide approximate probability estimates for the effects of interest. In this case, EMS equations are summed and/or subtracted to form appropriate error terms.

An Extension of Hopkins' Design

Hopkins (1984) demonstrated the application of the options above to the analysis of a repeated measures design in which persons were nested within treatment groups and crossed with items. The person and item factors were both considered as random factors. In this thesis, Hopkins' work on the incorporation of generalizability theory into statistical analyses is extended to a more complex data set in which persons are nested within gender and test items are nested within content domains and levels of cognitive complexity. In this case, the latter two factors are naturally fixed factors. Thus, the extension involves incorporating the structure of the dependent variable to more accurately reflect the way items are typically organized on a test or measuring instrument.

Because many educational measures are designed from tables of specifications with the numbers of items within each content area and level of the taxonomy preordained, they can be conceived of as being composed of a stratified random sample of items within each nesting variable. This type of structure for the dependent variable is commonly encountered in educational research. Unfortunately, the structure is also likely to be unbalanced, causing nonorthogonality among the facets considered and, consequently, further difficulties in analysis.

The data set used in this thesis consisted of the item scores obtained by a sample of students who wrote the June, 1989 British Columbia Algebra 12 departmental examination. Given the extant structure of the examination data, the use of a completely nested design involving the item, content domain, and cognitive complexity facets was not possible. Consequently, the post hoc analysis methods of model simplification and quasi F-ratios were employed, taking the nonorthogonality of the data into account. These analyses were conducted from the perspective of generalizability theory to demonstrate the advantages and disadvantages of its application to such situations.

Although further complexities could be introduced into the study by the introduction of additional independent variables, it was felt that a systematic increase in design complexity involving only the dependent variable was the best way to assess the practicability of Hopkins' proposed methods of data analysis for designs of varying complexity. A subsequent study is planned in

which Hopkins' proposed methods will be applied to a design in which additional independent variables are introduced. It is hoped that, with this three stage approach -- Hopkins (1984), the present study, and the future study -- alternative data analysis approaches which exist to handle non-orthogonal data in complex designs can be traced and assessed.

CHAPTER II

LITERATURE REVIEW

An Overview of Generalizability Theory

The ultimate purpose of scientific measurement is to acquire information about attributes or characteristics of objects or people. In all sciences, however, and especially in the social sciences, measurements are not perfect in that they contain some amount of error. The measurements obtained can be affected by a great number of factors related to variations in the object of measurement and in the conditions of measurement. These factors and their interactions can all potentially contribute to measurement error depending on the intended interpretation of the measurements.

Issues related to the estimation of measurement error in psychology and education have traditionally been explored within the framework of classical test theory. To obtain an estimate of the measurement error, one usually first computes a reliability coefficient and then uses it to compute the standard error of measurement (SEM). In classical test theory, there are four basic types of reliability corresponding to the measurement design used and, consequently, to the source of error to which the design is sensitive. Firstly, we may get an estimate of the stability of a particular measurement device over time or across raters, which is known as the coefficient of stability; SEMs calculated from this estimate reflect variation in measurements across time. We may also get an estimate of the stability of measurements when

different instruments are used. In this second case, the reliability coefficient is referred to as the coefficient of equivalence and the SEM based on this estimate reflects nonequivalence among instruments. For assessing the stability of measurements within a given instrument, a coefficient of internal consistency is computed. The SEM in this case reflects nonequivalence among the items. Lastly, there is the coefficient of stability and equivalence, which is an estimate of reliability across instruments and across time; the SEM based on this estimate reflects variations due to the time of measurement, the choice of instrument, and the interaction between these factors. Although the last case demonstrates the possibility of estimating the reliability of measurements when more than one dimension is varied, classical test theory does not allow us to differentiate among the three sources of error identified. The existence of a single error term in the classical test score model,

$$X_{ij} = \tau_i + \varepsilon_{ij},$$

inherently precludes any differentiation among multiple factors and their interactions as sources of error within the model.

A useful approach to the estimation of error variance and reliability coefficients when there are multiple sources of error involves the use of analysis of variance. Cronbach et al. (1972) comment that Fisher (1925)

> revolutionized statistical thinking with the concept of the factorial experiment in which the conditions of observation are classified in several respects.

> Investigators who adopt Fisher's line of thought must abandon the concept of undifferentiated error. The error term formerly seen as amorphous is now attributable to multiple sources, and a suitable experiment can estimate how much variation arises from each controllable source. (p. 1)

The use of analysis of variance to estimate measurement error variance and reliability has a long history that dates back to the work of Burt (1936). Most of the major formulas were worked out by researchers such as Hoyt (1941) and Lindquist (1953) long before Cronbach et al. (1963) published their first paper on generalizability theory.

Although Cronbach et al. (1972) did not provide any "new" formulas for estimating reliability per se, what they did do in their text was provide a framework by which one could systematically look at a measurement problem, decide upon what factors have potential relevance to the obtained measurements, and conduct a study by which the variance in observed scores attributable to each factor could be estimated. Once the sources of variance have been estimated, the next step is to decide which of these sources of variance contributes to universe score variance and which contributes to error variance. In generalizability theory, Cronbach et al. (1972) developed a systematic "way of thinking about reliability, which leads to procedures for choosing the reliability coefficient and/or error variance most appropriate for the situation at hand" (Crocker & Algina, 1986, p. 158).

The G-Study and Universes of Admissible Observations

Within the framework of generalizability theory, the first step in deciding on a procedure for measuring an attribute of an object of measurement, most frequently a person, is to specify the facets of measurement that might be relevant to the situation. A facet of measurement corresponds to a factor in the terminology of the ANOVA; it is a set of similar conditions of measurement. Any factor that can conceivably be a source of measurement error could be construed as a facet of measurement. The purpose of the G-study is to estimate the amount of variance in observed measurements that is attributable to each of the facets in the study, the objects of measurement, and the interactions among facets and objects of measurement.

Once the researcher defines the facets of measurement that are considered relevant to a study, the next decision to be made is how to define each facet. This involves the setting of reasonable boundaries to the populations within each facet based on the researchers perceptions and goals. The levels of a facet that fall within the boundaries defined by the researcher are known as the universe of admissible observations for that facet. For example, if a researcher wants to measure the degree of schizophrenic symptoms displayed by patients after they have been administered a new drug, he may decide that he will accept any clinical psychologist with more than three years of experience in dealing with schizophrenic patients as a rater whose judgement he will trust. In this case, "clinical psychologist with more than three

years of experience in dealing with schizophrenic patients" defines the universe of admissible observations for the rater (r) facet. Ratings made by people who do not fit this description would fall outside of the universe of admissible observations and would not be considered valid by the researcher. Other researchers may consider the boundaries of this universe to be too strict, however, and there is no constraint on their defining a different universe of admissible observations in another study. There is no absolute answer as to whose universe is the "correct" one, but the universes defined by various researchers should be kept in mind when interpreting the findings of different studies.

In the above study, the researcher might decide that it is important to look at the symptoms displayed by the patients at different times in order to assess the measurement error resulting from variations across time. This adds a second facet, time (t), to the study. It is likely that the researcher would accept the ratings of qualified raters regardless of the time of their observations. In this case, the raters facet would be considered crossed with the time facet in the universe of admissible observations; in the conventional notation of generalizability theory this would be denoted by r x t, where "x" is read "crossed with." Due to practical considerations, however, the researcher might find that it is necessary to have any single rater conduct all of his observations at a given time. In this case, each time of observation would be associated with a given set of raters. In this case, raters would

be considered to be nested within times; this would be denoted by r:t, where ":" is read as "nested within."

Clearly, if there is a universe of admissible observations based on defined boundaries of the facets of the study, there must also be a universe of people on whom the observations are admissible. Although this is true, the word "universe" is generally used to define the conditions of measurement (the facets) and the word "population" is used to define the objects of measurement in generalizability theory. This distinction is quite artificial, however, and it has been criticized as promoting the conceptualization of generalizability theory as a tool to build measures that better differentiate among persons tested (Cardinet, Tourneur, & Allal, 1981) rather than as "a powerful descriptive and analytic tool for other problems, where persons are not the central object of study" (Cardinet et al., 1976, p. 119). These authors stress the symmetry of generalizability theory: the fact that there is no logical basis for making a distinction between facets of a study and the objects of measurement when estimating statistical parameters. They argue that all factors in the ANOVA should be considered facets and suggest that depending on the goals of the study, facets relating to the objects of the study will constitute the face of differentiation and facets relating to the conditions of observation will constitute the face of instrumentation (Cardinet & Allal, 1983).

Once the researcher has defined the universe of admissible observations and the population on which those observations are to

be made, he is ready to collect data for the G-study and then estimate variance components by applying a factorial analysis of variance to the data collected. In the p x r x t design discussed here, unbiased variance components can be estimated through simple algebraic manipulation of the mean squares obtained from the ANOVA provided that the design is balanced (Brennan, 1983).

The Underlying Model for Generalizability Theory

The "fundamental equation", or underlying model, in generalizability theory is the decomposition of a collection of observed scores into a grand mean plus a number of score effects associated with each factor of interest in a G-study. A score effect is a parameter that reflects the unique contribution of a given level of a factor or of a given combination of levels of multiple factors to an observed score. For the p x r x t case, the equation decomposing observed scores into a grand mean and score effects would be:

$$X_{prt} = \mu + (\mu_p - \mu) + (\mu_r - \mu) + (\mu_t - \mu) + (\mu_{pr} - \mu_p - \mu_r + \mu) + (\mu_{pt} - \mu_p - \mu_t + \mu) + (\mu_{rt} - \mu_r - \mu_t + \mu) + [(\mu_{prt} - \mu_{pr} - \mu_{pt} - \mu_{rt} + \mu_p + \mu_r + \mu_t - \mu) + (X_{prt} - \mu_{prt})]. \quad (1)$$

In the above equation, X_{prt} represents the observed score for person p given by rater r at time t and μ_α represents the mean score for a given level of a factor or a given combination of levels of multiple factors, denoted by α. Thus μ_p, μ_r, and μ_t represent the person mean, rater mean, and time mean respectively. As

examples of the meanings of the interaction terms μ_{pr} represents the mean for all measurements of person p by rater r over time; μ_{prt} represents the mean score of person p given by rater r at time t. Because only one observation is used to estimate μ_{prt}, this estimate is completely confounded with the final term of the equation, the error term, as is indicated by the square brackets in Equation 1.

Equation (1) can be expressed more simply as:

$$X_{prt} = \mu + \mu_{p\sim} + \mu_{r\sim} + \mu_{t\sim} + \mu_{pr\sim} + \mu_{pt\sim} + \mu_{rt\sim} + [\mu_{prt\sim} + \varepsilon], \quad (2)$$

or

$$X_{prt} = \mu + \sum \mu_{\alpha\sim} + \varepsilon \quad (3)$$

where the symbol "~" denotes a score effect. Note that for higher order interaction terms, such as $\mu_{prt\sim}$, the calculation of score effects from mean scores can become quite complex. However, Brennan (1983) presents an algorithm for expressing score effects in terms of mean scores for designs of any complexity.

As a result of the definition of score effects in terms of mean scores (Equation 1), the expected value, E, of each effect is zero:

$$E_{p\sim} = E_{r\sim} = E_{t\sim} = E_{pr\sim} = E_{pt\sim} = E_{rt\sim} = E_{prt\sim} = E_{\varepsilon\sim} = 0. \quad (4)$$

Therefore, the equation for the variance for any given score effect, denoted by α, reduces to, $\sigma^2_\alpha = E\alpha^{-2}$. In the following section, an ANOVA based procedure for the estimation of variance that is applicable to balanced designs is presented.

An ANOVA Based Method of Estimating Variance Components

Assuming that each of the facets and the population in a study are of infinite size, the expected mean squares obtained from a p x r x t ANOVA design and derived using the procedures developed by Cornfield & Tukey (1956) are presented in Table 1. Note that $\sigma^2(prt,e)$ is presented as one component rather than as two separate components. This is a convention utilized by Cronbach et al. (1972) to indicate the complete confounding of the highest level interaction component with other unidentified

Table 1

Expected Values of Mean Squares Based on the Cornfield & Tukey (1956) Algorithms

$$E[MS(p)] = \sigma^2(prt,e) + n_t\sigma^2(pr) + n_r\sigma^2(pt) + n_r n_t\sigma^2(p)$$

$$E[MS(r)] = \sigma^2(prt,e) + n_t\sigma^2(pr) + n_p\sigma^2(rt) + n_p n_t\sigma^2(r)$$

$$E[MS(t)] = \sigma^2(prt,e) + n_r\sigma^2(pt) + n_p\sigma^2(rt) + n_p n_r\sigma^2(t)$$

$$E[MS(pr)] = \sigma^2(prt,e) + n_t\sigma^2(pr)$$

$$E[MS(pt)] = \sigma^2(prt,e) + n_r\sigma^2(pt)$$

$$E[MS(rt)] = \sigma^2(prt,e) + n_p\sigma^2(rt)$$

$$E[MS(prt)] = \sigma^2(prt,e) = \sigma^2(prt)$$

sources of error and random measurement error. Others, though, represent this component using only the term $\sigma^2(prt)$, and this is the convention adopted in the remainder of this thesis.

If the expected mean squares from the formulas in Table 1 are replaced by the observed mean squares from the ANOVA, unbiased estimates of variance components can be obtained by the application of linear algebra to these equations to isolate individual variance components. The results of this manipulation for the above design are presented in Table 2.

Table 2

Estimated Variance Components for a p x r x t G-Study Design

$\hat{\sigma}^2_p = (MS_p - MS_{pr} - MS_{pt} + MS_{prt})/n_r n_t$
$\hat{\sigma}^2_r = (MS_r - MS_{pr} - MS_{rt} + MS_{prt})/n_p n_t$
$\hat{\sigma}^2_t = (MS_t - MS_{pt} - MS_{rt} + MS_{prt})/n_p n_r$
$\hat{\sigma}^2_{pr} = (MS_{pr} - MS_{prt})/n_t$
$\hat{\sigma}^2_{pt} = (MS_{pt} - MS_{prt})/n_r$
$\hat{\sigma}^2_{rt} = (MS_{rt} - MS_{prt})/n_p$
$\hat{\sigma}^2_{prt} = MS_{prt}$

Application of Variance Components to the Design of Measurement Procedures

The ultimate goal of a G-study is to obtain accurate estimates of variance components associated with the universe of

admissible observations. One important use of these estimates is to design efficient measurement procedures that will be reliable for making inferences about the objects of measurement in a D-study. For instance, if one finds that an unacceptable amount of measurement error is being introduced into a study from a given source, it may be decided that more measurements should be taken on that facet in order to reduce the total error variance to a reasonable level. Continuing with the p x r x t example introduced in the previous section, if the variance component for raters indicated that raters contributed a sizable amount of variance to the ratings of the symptoms displayed by schizophrenics, a researcher comparing the effects of different antipsychotic drugs might decide to have two or three raters examine each patient in order to reduce this source of variance in his study.

The D-Study and Universes of Generalization

D-studies emphasize the estimation, use, and interpretation of variance components for decision making. Although the distinction between G-studies and D-studies is important conceptually, in many cases the same data are used for both. Central to a D-study is the universe of generalization, which is the universe to which the decision maker intends to generalize the measurement results. Although the definition of the universe of generalization is dependent on the researcher's goals, it is logically necessary that it be a subset of the universe of admissible observations. If G- and D-studies use the same design and define the facets in the same way as being random or fixed,

the universe of admissible observations and universe of generalization will be identical. However, even with the same data set, the two universes need not be co-terminus. This situation commonly occurs when a facet that is defined as random in the G-study is defined as fixed in the D-study. Generalizability theory allows the consideration of many possible universes of generalization, providing that each is a subset of its corresponding universe of admissible observations. In fact, Hopkins (1983) advocates an approach to data analysis for certain cases such that the universe of generalization is initially restricted, and systematically broadened until it coincides with the universe of admissible observations.

The number of levels of each variable (i.e., the sample sizes) need not be the same in a D-study as in a G-study. In fact, if one has used the results of a G-study as a guide to maximizing the efficiency of the D-study, it is unlikely that the sample sizes will be identical. In order to distinguish D-study sample sizes from G-study sample sizes, they are generally denoted with a prime. Thus, for the above example, D-study sample sizes would be denoted by n_p', n_r', and n_t'.

The Structure of D-Study Designs

D-study designs may also have a different structure than the G-study. To begin with, in a D-study one is usually not interested in the variance components associated with the facets in the face of instrumentation per se. Instead, it is with the error variance associated with mean scores over the sets of conditions of these

facets that the researcher is generally concerned. To illustrate that it is the mean scores in the facets of instrumentation that we work with in a D-study, these facets are generally denoted by upper case letters. Thus, continuing with the example introduced above, a likely D-study design would be p x R x T.

Another aspect of the structure of a D-study design that may differ from that of a G-study is the relationship among the facets with respect to nesting and crossing. The ideal case for a G-study is to have a fully crossed design because this allows the estimation of all possible variance components for any subsequent D-study design, whether there be nesting or not. The effect of nesting is to confound the variance associated with the nested variables with that due to the interaction with the variables within which they are nested. Consequently, one can estimate all of the variance components for a nested design from a crossed design, but not vice versa. As is pointed out later, there are occasions in which the confounding of effects associated with nesting may actually be desirable in a D-study. Of course, in many situations the nesting or crossing of variables in both G- and D-studies is more dependent on practical considerations than on ideal cases. Furthermore, the relationships among some variables are such that nesting is the natural state and it would be inconceivable to have them crossed in any design.

Estimation of Variance Components in D-Studies

The effect of estimating D-study variance components based on mean scores on the facets of instrumentation is directly

analogous to the progression from the variance of a population of scores to the variance of sample means,

$$\sigma^2_{\bar{X}} = \sigma^2_x / n_x.$$

For the case where the variance component of interest, α, contains more than one facet of instrumentation that is to be averaged, one simply divides σ^2_α from the G-study by the sample sizes for each facet of instrumentation contained in α. For example, the D-study variance component for $\sigma^2(pRT)$ would be $\sigma^2(prt)/n'_r n'_t$.

Fixed Versus Random Facets

The facets that make up the universe of generalization can be conceived of as being either random or fixed and often the decision as to how they are defined is dependent on how the researcher wishes to interpret the results of a study. If one would like to generalize from the specific levels of a facet that were applied in a study to a larger population of possible levels of that facet, then it is likely that the facet in question should be characterized as random. On the other hand, if one has already sampled all of the conditions from a population or is not interested in generalizing past those that were sampled in a study, then the facet in question would best be characterized as fixed.

Universe Scores

Each object of measurement in a D-study can be assigned a universe score. A universe score is defined as the mean score for the object of measurement over all of the conditions in a given

universe of generalization. Thus, it is very similar in nature to the concept of the true score from classical test theory. The universe score for a given patient in the above study would be,

$$\mu_p = E_r E_t(X_{prt|p}),$$

the expected value of the observed scores for the person of interest taken over the conditions of the rater facet and the time facet. Mean "scores" can be defined in a similar manner for conditions of measurement (ie., raters, times) as well, but in generalizability theory the term universe score is reserved for objects of measurement.

Universe Score Variance

One reason for differentiating between universe scores and true scores is that the variance of universe scores is dependent upon the universe of generalization defined by the researcher. Generalizability theory allows the universe score variance to change depending upon how the universe scores are interpreted. If any facets in the universe of generalization are considered fixed as defined in the next section, then the variance due to interaction of these facets with the objects of measurement contributes to universe score variance. If they are considered random facets, then their variance contributes to error variance.

Absolute and Relative Measurement

Measurements can be divided into two general classes, absolute measures and relative measures. Absolute measures assign quantitative scores on a given dimension to the object of measurement that are considered to be invariant with respect to

external factors. Relative measures, on the other hand, assign scores to the object of measurement that allow one to compare it to 'similar' objects of measurement on a given dimension. Thus, quantifying relative standing in comparison to the norm is the focus of relative measurement. In terms of educational measurement, criterion-referenced score interpretations represent absolute measurement and normative score interpretations represent relative measurement.

Naturally, both types of measurement have errors associated with them. Based on the above definition, it is clear that for the case of absolute measurements, any variation in person scores associated with variations in the conditions of measurement is indicative of absolute measurement error. In the example of the schizophrenia study, this type of error can be defined as $X_{prt} - \mu_p$, the difference between a score assigned to a given person and the person's average score over the theoretically infinite number of possible measurement conditions.

For the case of relative measurement, the sources of error are more limited in scope. Only those sources of variation that result in changes in the rankings of the objects of measurement in a given group are sources of relative error; variation in measurement conditions that affect all objects of measurement equally are not sources of relative error. Using the schizophrenia study example, these constant score effects include the effects $\mu_{r\sim}$, $\mu_{t\sim}$, and $\mu_{rt\sim}$. By subtracting these constant score effects from the equation for absolute error, relative error can be defined

as $(X_{prt} - \mu_p) - \mu_{r\sim} - \mu_{t\sim} - \mu_{rt\sim}$. Alternatively, by substituting these score effects with their counterparts from Equation 1 and simplifying the resulting equation, relative error can be defined as $(X_{prt} - \mu_p) - (\mu_{rt} - \mu)$.

Absolute and Relative Error Variance

Absolute error variance is denoted in generalizability theory as $\sigma^2(\Delta)$. This is the variance of the difference between observed scores and the universe score for each object of measurement. Because any deviation of observed scores from universe scores is undesirable when absolute error is of concern, all sources of variance in the D-study that do not constitute universe score variance contribute to absolute error variance. Therefore, all of the variance estimates involving random facets from the face of instrumentation can be summed to obtain an estimate of $\sigma^2(\Delta)$.

Relative error variance is denoted as $\sigma^2(\delta)$. The magnitude of $\sigma^2(\delta)$ is dependent upon differences in the rank order among the objects of measurement. In a fully crossed design, $\sigma^2(\delta)$ can be estimated by summing the variance components for interactions between random facets of instrumentation and the objects of measurement.

Generalizability and Dependability Coefficients

One simple method which can be used to assess the magnitude of universe score variance for a given application of a measurement procedure is to compare it to error variance using something analogous to a signal-noise ratio (Brennan, 1983). Although this would be quite a reasonable approach, it is much

more common to construct something analogous to a reliability coefficient that is known as a generalizability coefficient. If absolute error variance is of concern to a researcher, the appropriate generalizability estimate to use is

$$\hat{\rho} = \hat{\sigma}^2_U/[\hat{\sigma}^2_U + \hat{\sigma}^2_\Delta],$$

where $\hat{\sigma}^2_U$ is the estimated universe score variance and $\hat{\sigma}^2_\Delta$ is the estimated absolute error variance. When relative decisions are of interest, the appropriate form for a generalizability coefficient is

$$\hat{\rho} = \hat{\sigma}^2_U/[\hat{\sigma}^2_U + \hat{\sigma}^2_\delta],$$

where $\hat{\sigma}^2_U$ represents estimated universe score variance and $\hat{\sigma}^2_\delta$ represents estimated relative error variance.

For the case of criterion-referenced measurement in which a given cutoff score is used to segregate a group, Brennan and Kane (1977) developed an index of dependability, the phi coefficient, which is defined as,

$$\hat{\Phi} = [\hat{\sigma}^2_U + (\hat{\mu} - \lambda)^2]/[\hat{\sigma}^2_U + (\hat{\mu} - \lambda)^2 + \hat{\sigma}^2_\Delta],$$

where $\hat{\mu}$ is an estimate of the mean and λ is the cutoff score for the instrument in question. For the case where $\lambda = \mu$, the phi coefficient is equal to the generalizability coefficient for absolute decisions.

Although they are very similar in form, reliability coefficients and generalizability coefficients are not identical.

Generalizability coefficients can be "tailor-made" to fit a given universe of generalization, so they allow more flexibility than reliability coefficients. As a result of this congruence between the generalizability coefficient and the desired universe of generalization, they also tend to be better estimators of the degree of measurement error. As they tend to focus on one source of error at a time, Hopkins (1984) argues that classical reliability coefficients ordinarily underestimate the degree of measurement error in the appropriate universe of generalization. This assumes, of course, that the appropriate universe of generalization includes more facets than those that are treated in the calculation of any single reliability coefficient.

Issues in the Estimation of Variance Components in Generalizability Theory

Difficulties in the estimation of variance components have been described as the "Achilles heel" of generalizability theory and of sampling theories in general (Shavelson & Webb, 1981). These difficulties arise from two main sources, (a) sampling variability that can lead to negative variance estimates and (b) difficulties in estimation from unbalanced designs. Although these problems can occur in all studies utilizing sampling procedures, the emphasis placed on the estimation of variance components in generalizability theory makes them particularly relevant in this context.

Sampling Variability and Negative Variance Estimates

Estimates of variance components have their own sampling variability and these estimates tend to be relatively unstable, especially for small sample sizes (Cronbach et al., 1972; Marcoulides, 1990; Shavelson & Webb, 1981). It should be noted that because variance components associated with each facet in a study are important in generalizability theory, sample sizes must be adequate not only on the person variable, but also on all other random variables in question.

This need for adequate sample sizes on all random variables is rarely mentioned explicitly in expositions on the estimation of variance components and this issue is given even less attention in papers which serve to "popularize" generalizability theory (eg., Malloy & Kenny, 1986; Shavelson, Webb, & Rowley, 1989). Even in a paper that has a reasonable amount of discussion about the variability of variance estimates for small sample sizes (Shavelson & Webb, 1981), an example is provided of a study with a three way crossed design in which 20 subjects were crossed with only two levels of each of the facets in the face of instrumentation (Gleser, Green, & Winget, 1978). As a result of the lack of stress placed on the need for adequate sample sizes in the current literature, it is this author's opinion that many researchers who apply generalizability theory to their data may not be aware of it. Smith (1978) argues that "estimates of variance components in multifacet generalizability studies contain sizable error and may be so unreliable as to be useless unless the

design utilizes a substantial number of levels of each facet. Little or nothing will be learned from small sample generalizability studies" (p 342).

Smith (1978, 1982) suggests that the minimum number of levels of a facet in a generalizability study for an adequate estimate of variance should be 10. This is a rough guideline, however, because sampling errors of variance estimates are dependent upon complex relationships between the sample sizes and the design configuration used in a given study (Smith, 1978). As a general rule, the more variance components incorporated into a formula for expected mean squares of a given component, the greater the probability of serious error in estimation (Smith, 1978). This is due to the possibility of dispersion effects, a term introduced by Leone and Nelson (1966) to describe the effect of errors lower in the hierarchy progressively distorting variance component estimates higher in the design.

One consequence of the instability of variance estimates is that as a true variance component approaches zero, the chance that ANOVA-based methods of estimation will result in negative estimates approaches fifty percent (Marcoulides, 1990). The likelihood of negative variance estimates also increases when the other components contained in the formula for the expected mean squares are relatively large in comparison to the component of interest (Smith, 1982). Although several methods have been proposed to deal with the existence of negative variance components, none of them are really adequate. Cronbach et al.

(1972) recommended setting negative variance estimates to zero and applying the zero estimate to any mean squares equations involving the variance component in question. This strategy produces biased estimates, however (Brennan, 1983). Brennan (1983) recommends that when focusing on a particular variance component, a negative estimate should be set to zero, but that negative estimates should be retained for the estimation of other variance components. This results in the aesthetically displeasing situation of using estimates that are logically impossible in further calculations. The most appealing solution to the problem of negative estimates is the use of techniques such as maximum likelihood or Bayesian analyses that inherently do not allow negative estimates to be calculated so that post hoc manipulation of the estimates is not necessary.

Unbalanced Designs

Another complicating factor in the estimation of variance components to be used in generalizability theory involves the problems created by unbalanced designs. In the preceding sections it has been implicitly assumed that the G- and D-studies have employed the ideal case of balanced designs. Unfortunately, this is not always the case for a variety of practical reasons. Although the calculation of variance components from mean square equations is relatively straightforward for the case of balanced designs, unbalanced designs complicate the necessary procedures considerably.

For balanced designs, ANOVA estimators have the desirable properties of (1) unbiasedness (assuming negative variance estimates have not been set to zero), (2) minimum variance among all unbiased estimators that are quadratic functions of the observations, and, (3) given normality, minimum variance among all unbiased estimators (Brennan, Jarjoura, & Deaton, 1980; Swallow & Monahan, 1984). With unbalanced data, however, properties (1) and (2) are lost. The lack of bias is also lost if negative estimates have been set to zero (Brennan et al., 1980; Swallow & Monahan, 1984).

Choosing a Method of Estimating Variance for Unbalanced Designs

A number of alternative methods have been developed to estimate variance components for unbalanced designs, but not all of them are appropriate or practical for a given G- or D-study design. Different methods have their own strengths and weaknesses, so choosing the appropriate one involves consideration of several factors. These factors fall into three general categories: (a) properties and appropriateness of the method, (b) accuracy of the variance estimates obtained, and (c) the computing resources required. These considerations will be discussed in the following sections.

Properties and Appropriateness of the Method

The appropriateness of several methods for estimating variance components in generalizability theory was assessed by Brennan et al. (1980, pp. 36 - 44) and this section follows their

discussion closely. They divided the various available techniques into two categories, ANOVA-like procedures and general approaches.

ANOVA-like Procedures

The procedures discussed in this section all utilize the technique known as the method of moments for estimating variance components. These procedures are variations on the basic ANOVA method described by Cornfield and Tukey (1956) and illustrated earlier (see Tables 1 and 2). They all involve the same basic steps of:

(a) calculating specified quadratic functions (forms) of the observations;

(b) determining the expected values of the quadratic functions in terms of the variance components in the model; and

(c) solving the set of linear equations resulting from equating the quadratic functions with their expected values.

These procedures require no distributional assumptions and all provide unbiased estimates providing the appropriate model is used and negative estimates are not altered. When applied to balanced designs, they all reduce to the basic ANOVA procedure and produce the same results. Although these methods are quite similar in nature, Brennan et al. (1980) argue that in the case of unbalanced designs, they can result in quite different estimates. The structure of the data set in question is the most important consideration in the decision as to what method is most

appropriate. Although a number of estimation procedures based on the method of moments exist (eg., Type 1 to Type 4 analyses), the following discussion is restricted to the well known procedures proposed by Henderson (1953).

Henderson's Method I. This is the first of three methods of estimating variance proposed by Henderson. The quadratic forms used for this method are similar to the sum of squares terms used with balanced designs, but they are adjusted to reflect the unbalanced nature of the design. Although the algebra required to calculate expected mean squares for this method can be tedious, Hartley's (1967) method of synthesis can be used to simplify computation. Synthesis can also be used to obtain estimates of the variance of the variance component estimates under the assumption of normality and therefore confidence intervals can be calculated. Given the large potential for error in estimates of variance components, the ability to obtain confidence intervals can be of significant value in guiding the interpretation of results of a generalizability study. However, Method I is designed for the estimation of variance components in a completely random model and is strictly applicable only to this case.

Henderson's Method II. This procedure was developed for application to the mixed ANOVA model. The observations are adjusted by estimating the fixed effects, partitioning them out and then applying Method I to the adjusted observations. This method does not allow the estimation of variance due to interactions

between fixed and random effects, which effectively makes it inapplicable to generalizability theory.

Henderson's Method III. This procedure can be applied to any mixed model to provide unbiased estimates of variance. It is essentially a regression analysis approach and is often referred to as the fitting constants method because specified sub-models are fitted in order, with their corresponding variance components being partitioned in the order specified. A consequence of this procedure is that a change in the specified order of fitting can result in a change in the estimates of variance components. The problem of justifying a given order of fitting increases as designs become more complex and contain more terms.

In many cases, the order of fitting is quite arbitrary and the problem of specifying a logically justifiable order of fitting increases with the complexity of the design (Brennan et al., 1980). In some situations, however, a convincing argument can be made for choosing a specific order of fitting. Brennan et al. comment that many solutions have been suggested for cases in which a specific order of fitting is not obvious, including the use of all possible logical orderings. Synthesis (Hartley, 1967) can also be applied to arrive at expected mean squares for this method.

General Approaches

In recent years, a number of approaches to the estimation of variance components in the unbalanced situation that apply for any design have been developed. "In contrast to the above ANOVA-like procedures, which might be considered ad hoc solutions to

unbalanced estimation, the newer approaches focus on providing a theoretical framework for the general variance components model and on optimal properties of the estimates (cf. Rao, 1971a,b; Harville 1977)" (Brennan et al., 1980, p.41) Two classes of estimators with desirable properties are maximum likelihood estimators and minimum norm quadratic unbiased estimators. Although there are notable differences between them, these two classes have strong links (Swallow & Monahan, 1984).

Bayesian techniques also deserve mention, but they are based on assumptions of normality (Searle, 1971) and they also lead to difficulty in the treatment of unbalanced data (Harville, 1977). In spite of this, several authors have argued for the potential usefulness of Bayesian techniques of estimating variance components in generalizability theory and suggest that they should be further developed (eg, Cronbach et al., 1972; Novick, 1975; Shavelson & Webb, 1981).

Maximum likelihood estimation. There are two types of maximum likelihood estimation commonly found in the literature. In unrestricted maximum likelihood estimation (ML), the values that maximize the full likelihood function are estimated. In restricted maximum likelihood (REML) estimation, on the other hand, it is only the portion of the likelihood function that is free of fixed effects that is maximized (Swallow & Monahan, 1984). In both cases, the variance component estimates for the random effects are invariant with respect to the values of the fixed effects; this is known as translation invariance. Of course, both

ML and REML also have the desirable properties of always producing nonnegative variance estimates due to the explicit definition of the parameter space as non-negative. However, this restriction on the parameter space results in biased estimates.

Other desirable properties of the ML and REML methods are that they are functions of sufficient statistics, they are asymptotically normal and efficient, and they provide asymptotic variances and covariances of the estimates (Brennan et al., 1980; Harville, 1977). These properties depend upon normality assumptions, however, which are likely to be violated in many studies. This is not a major drawback, however, especially considering the fact that ML procedures share a strong relationship to the MINQUE procedures to be discussed in the next section (Searle, 1979, cited in Brennan et al., 1980).

One problem with ML estimation of variance components is that the loss of degrees of freedom due to the estimation of fixed effects is not taken into account (Harville, 1977). As a result of this, the variance component estimates obtained by solving ML equations tend to be biased in a downward direction. Consequently, they tend not to coincide with those obtained by ANOVA methods, even in the case of balanced designs. These problems are eliminated in the case of REML estimation, because REML estimates are adjusted for the degrees of freedom lost from the estimation of fixed effects. A pleasing result of this is that REML estimates are identical to ANOVA estimates for balanced data (Brennan et al., 1980; Harville, 1977).

Minimum norm quadratic unbiased estimation. Minimum norm quadratic unbiased estimation (MINQUE) is a general technique developed by LaMotte (1973) and Rao (1971a,b) that provides a general solution to the problem of the estimation of variance components. MINQUE estimates have the characteristic of having local minimum mean squares of error; in other words, "they are locally best when attention is restricted to estimators satisfying various conditions" (Harville, 1977, p. 333). These conditions are that the estimator minimize a Euclidean norm, be a quadratic form of the observations, and be unbiased. No distributional assumptions are required for the calculation of MINQUEs.

Minimum variance quadratic unbiased estimators (MIVQUEs) are a subtype of MINQUEs for which normality assumptions have been invoked and which have the property of translation invariance (Harville, 1977). To calculate MIVQUEs, one must supply a priori values for the variance components and these values are used in combination with the data to arrive at the final estimates. The a priori values used in the calculation of MIVQUEs usually come from one of two sources: (a) the output from another variance estimation procedure or (b) default values of one for error variance and zero for all other variance components in the model. Unfortunately, the property of minimum variance among estimators in general holds for MIVQUEs only when each a priori value equals the true value of its corresponding variance component (Swallow & Monahan, 1984).

Harville notes that both MINQUEs and MIVQUEs are not estimators at all in the strictest sense because they can produce estimates that violate the restrictions on the parameter space. Under normality, MINQUE and MIVQUE equations are identical and with balanced data they produce identical results to ANOVA based and REML methods (Swallow & Monahan, 1984). Furthermore, the formula for the MIVQUE is identical to that for REML with the a priori weights in MIVQUE taking the place of variance estimates arrived at through iteration in REML (Harville, 1977; Swallow & Monahan, 1984). Consequently, if the obtained variance estimates from the MIVQUE are used as the starting weights for a new run, and this process is iterated until convergence, the results will be equivalent to the REML solution (Searle, 1987).

Accuracy of Estimates

Although ANOVA procedures require no assumptions about the distributional form of any effects (Brennan et al., 1980), deviations from normality result in an increase in the standard error of the variance estimates obtained (Swallow & Monahan, 1984; Muthen, 1983, cited in Marcoulides, 1990). In contrast, although maximum likelihood approaches are derived on the basis of normality, Marcoulides (1990) demonstrated that REML estimates are consistently as accurate or more accurate than ANOVA based estimates for a number of balanced and unbalanced designs of varying complexity and involving a varying amount of deviation from normality. He found that REML estimates tended to be closer to the true parameters for the distributions and also had

less sampling variability than ANOVA based estimates. In spite of this conclusion, he cautions that although "the sampling variability of REML estimates is smaller than that for analysis of variance, it is still quite sizable and would yield considerably wide confidence intervals of variance components" (Marcoulides, 1990, p. 385).

Swallow and Monahan (1984) performed a Monte Carlo comparison of ANOVA, ML, REML, and MIVQUE estimators of variance components for a one-way random model with unbalanced data. They varied the between group variance (σ^2_a) from a low of zero to a high of five and maintained the within group variance (σ^2_ε) at one. Because Swallow & Monahan enforced nonnegativity for all of the estimators, the property of being unbiased was lost for the ANOVA and MIVQUE estimators, so all of the methods used produced biased results. For the ML estimator, the downward bias in the estimates resulting from its failure to take the loss of degrees of freedom associated with estimation of fixed effects became quite pronounced as σ^2_a became larger, but for smaller values of σ^2_a this effect was counterbalanced by the upward bias resulting from the enforcement of nonnegativity. The ANOVA, MIVQUE, and REML estimators differed little in bias, but the expected upward effect due to enforcing nonnegativity did increase as σ^2_a approached zero; biases for these estimators for σ^2_a were negligible, however.

In the Swallow and Monahan (1984) study, two variations of the MIVQUE program were evaluated. The first variation, designated MIVQUE(0), is the default condition for the SAS

PROCEDURE VARCOMP (SAS Institute, 1988), in which a priori estimates of $\sigma^2_a = 0$ and $\sigma^2_\varepsilon = 1$ are used. The second variation, designated MIVQUE(A), uses ANOVA results to generate a priori estimates. Swallow and Monahan found that although MIVQUE(0) is widely used, it is a poor estimator for σ^2_a and a "terrible" estimator for σ^2_ε for the case in which $\sigma^2_a/\sigma^2_\varepsilon \geq 1$. Based on this result, they comment that "MIVQUE(0) is a dangerous default and should be used only when one is confident that $\sigma^2_a = 0$" (p. 54). They also conclude that unless the data are severely unbalanced and $\sigma^2_a/\sigma^2_\varepsilon > 1$, ANOVA estimators are adequate. MSEs were found to be similar among the four methods for all other conditions, which suggests that for most instances of the one-facet case, factors other than accuracy can take precedence in the choice of variance estimation method.

Computing Resources Required

Chastain and Willson (1986) performed several generalizability analyses on the data from the WAIS-R standardization sample (n = 1880) using the four estimation methods available in the SAS procedure VARCOMP: (a) TYPE 1, (b) MIVQUE(0), (c) ML, and (d) REML. The TYPE 1 method computes the Type 1 sum of squares for each effect, equates each mean square involving only random effects to its expected value, and solves the resulting system of equations (Gaylor, Lucas, & Anderson, 1970). This method is equivalent to Henderson's Method III.

Chastain & Willson (1986) divided the standardization sample for the WAIS-R into four groups of 470 by random

assignment and used the above programs to estimate variance components for three unbalanced crossed designs: (a) a two-factor subject by item design, (b) a three-factor subject by item by age design, and (c) a four-factor subject by item by age by gender design.

Although the results of the studies presented in the preceding section suggest that REML is the most consistently accurate method of variance component estimation, the data presented in Table 3 indicate that the price for this accuracy is paid for in terms of computing time required to reach convergence. Even after 360 seconds of c.p.u. time, REML still had not reached convergence. These results are consistent with those reported by Bell (1985), who makes the case that because of its relative efficiency and ease of use, MIVQUE(0) is the most useful software currently available for generalizability analyses. This position is difficult to reconcile with that of Swallow and Monahan (1984) who argue that MIVQUE(0) is a "dangerous default." Clearly, there is a need for more research into accurate and computationally efficient algorithms for estimation of variance components in unbalanced designs. Among the four methods presented in Table 3, however, it appears that the Type 1 method offers the most reasonable compromise between accuracy of estimates and demands on computing resources.

The Fixed-Effect Fallacy

In a seminal paper titled "Generalizing to a Language Population," Coleman (1964) voiced his concern over the

Table 3

Average C.P.U. Time and Memory Requirements for Estimation of Variance Components for Unbalanced Designs Using Various Methods

Method	Number of Factors	Average C.P.U. Time (seconds)	Average Memory Requirement
TYPE 1	2	33.6	620 K
	3	56.6	1270 K
	4	91.3	2050 K
MIVQUE(0)	2	2.1	620 K
	3	3.0	1120 K
	4	3.7	2000 K
REML	2	over 360	3000 K
	3	over 360	3000 K
	4	over 360	3000 K
ML	2	over 360	3000 K
	3	over 360	3000 K
	4	over 360	3000 K

incongruity between statistical analyses and the conclusions drawn from the results of these analyses in language research. He argued that,

> Many studies of verbal behavior have little scientific point if their conclusions have to be restricted to the specific language materials that were used in the experiment. It has not been customary, however, to

> perform significance tests that permit generalization beyond these specific materials, and thus there is little statistical evidence that such studies could be successfully replicated if a single sample of language materials were used. (p. 219)

Nine years later, Clark (1973) pointed out that Coleman's criticisms had "been all but totally ignored ever since" (p. 335). Using a classic series of studies in the field of semantic memory as an example, Clark demonstrated that with appropriate statistical analyses, these studies "provide no reliable evidence for most of the main conclusions drawn from them" (p. 335). He coined the term *language-as-fixed-effect fallacy* to describe the problem.

Based on the number of citations of Coleman's and Clark's papers in various journals, Malgady, Amato, and Huck (1979) provided evidence that although the language-as-fixed-effect fallacy was beginning to be acknowledged by researchers in the area of basic psychology, little attention was being paid to this threat to external validity by educational researchers. They found that more than 90% of the studies reviewed in the 1976 volume of the *Journal of Educational Psychology* failed to use appropriate statistics to demonstrate that their conclusions generalized to a meaningful language population. Malgady et al. also pointed out that the language-as-fixed-effect fallacy should be considered a special case of a more general statistical error of treating random effects as fixed effects due to failure to identify relevant sampling variables. They specifically noted the relevance of the

fixed-effect fallacy in the context of educational measurement, stating that "items on an intelligence test may be considered a sample from the universe of items measuring intellectual ability; summing item scores to create a total IQ score is therefore tantamount to the fixed-effect fallacy" (p. 85). They also made a reference to the utility of the distinction in generalizability theory between the universe of admissible observations and the universe of generalization in providing a clear framework for the decision as to whether conditions should best be considered fixed or random.

Hopkins (1984) focused on cases in which cognitive or affective measures serve as the dependent variable in a study and made identical arguments regarding the incongruity between analysis and inference to those of Coleman and Clark. He went beyond this, however, to propose a systematic integration of generalizability theory and experimental design. In Hopkins' proposed method, items on tests or inventories are treated as levels of a random factor in the ANOVA design, thus circumventing the fixed-effect fallacy. His method has the additional advantage of allowing the calculation of a generalizability coefficient for the dependent variable being assessed using the same data. However, although generalizability theory does provide conceptual clarity to the situation, complications are introduced into the statistical analysis of a study by the inclusion of multiple random effects.

Methodological Solutions to the Fixed-effect Fallacy

Experimental Design

The most appealing way by which the fixed-effect fallacy can be avoided is to design studies such that appropriate inferential statistics can be applied to the data as it stands without necessitating unusual statistical manipulations. In the discussion to follow, it is assumed that the inferential statistic to be employed is an F-test. As a general rule, if the random variables are confounded such that the mean square term for the effect of interest (the treatment effect) differs by the term for the residual mean square by only the variance component for the treatment effect, a simple F-ratio can be calculated which will be associated with the appropriate probability value. To examine this more closely, we can examine the designs presented in Hopkins (1984).

Hopkins presented analyses of the results of a ten-item examination administered to subjects nested within two treatment groups. Traditionally, this design would be analyzed as a one-factor ANOVA with treatments (T) and persons nested within treatments (P:T) as the two sources of variation. In this design, items are implicitly treated as a fixed facet. To make this underlying assumption clear, if the item factor (I) is included in this design, but is still considered fixed, we arrive at the repeated measures design presented in Table 4. In this design, note that there is an appropriate error term for the treatment (T) effect, namely persons nested within treatment groups (P:T). However,

due to the treatment of items as fixed, generalization can only be justifiably made to the particular sample of items employed.

Table 4

ANOVA Table for the P:T x I Design (I Fixed)

SV	DF	EMS
T	J - 1	$K\sigma^2_{p:t} + nK\sigma^2_t$
I	K - 1	$\sigma^2_{ip:t} + nJ\sigma^2_i$
P:T	J(n - 1)	$K\sigma^2_{p:t}$
TI	(J - 1)(K - 1)	$\sigma^2_{ip:t} + n\sigma^2_{ti}$
IP:T	J(K - 1)(n - 1)	$\sigma^2_{ip:t}$

Items can defined as a random factor in the P:T x I design to allow generalization to other similar items, but this results in an alteration of the EMS equations such that there is no longer an appropriate error term with which to test the treatment effect (see Table 5). Thus, model simplification or quasi F-ratios, with their associated difficulties, would have to be utilized to test the treatment effect for this case. These techniques will be discussed more fully in the next section.

One way to avoid the problems associated with model simplification and quasi F-ratios is to design a study such that it is possible to conduct an F-test on the effect of interest with an appropriate error term. This can be accomplished by completely

Table 5

ANOVA Table for the P:T x I Design (I Random)

SV	DF	EMS
T	J - 1	$\sigma^2_{ip:t} + n\sigma^2_{ti} + K\sigma^2_{p:t} + nK\sigma^2_t$
I	K - 1	$\sigma^2_{ip:t} + nJ\sigma^2_i$
P:T	J(n - 1)	$\sigma^2_{ip:t} + K\sigma^2_{p:t}$
TI	(J - 1)(K - 1)	$\sigma^2_{ip:t} + n\sigma^2_{ti}$
IP:T	J(K - 1)(n - 1)	$\sigma^2_{ip:t}$

confounding the variables in the study to allow simultaneous generalization to multiple populations (Coleman, 1964; Hopkins, 1984; Malgady et al., 1979). In the above example, the populations that we would like to generalize the results to are persons and the items that they respond to. For this example, the appropriate design to use that allows one to directly test the treatment effect would be a hierarchical one with I:P:T. The ANOVA table for this design is reproduced in Table 6.

Table 6

ANOVA Table for the I:P:T Design

T	J - 1	$\sigma^2_{ip:t} + K\sigma^2_{p:t} + nK\sigma^2_t$
P:T	J(n - 1)	$\sigma^2_{ip:t} + K\sigma^2_{p:t}$
I:P:T	nJ(K - 1)	$\sigma^2_{ip:t}$

Although the technique of confounding of variables is advantageous with respect to ease of statistical analysis for the treatment effect, it has certain drawbacks. Firstly, the confounding of item variance and person variance in this design will result in an increase in the magnitude of $\sigma^2_{p:t}$ and consequently some increase in $MS_{p:T}$, and this will not be completely offset by the concomitant increase in MS_T. Therefore, the hierarchical design results in a decrease in power and a greater probability of a Type II error than the factorial design. A second concern is that the random variables have to be sampled from their respective populations and this could easily result in a huge amount of labour on the researcher's part, depending upon the difficulty involved in obtaining a large sample of the variables in question (Malgady et al., 1979). When one considers the previously discussed need for the large number of levels of each random facet for stable estimates of variance, the amount of labour required could become staggering. This labour can be reduced greatly if an item bank already exists.

A final problem with this type of design is that it is only practical if one only wants to test hypotheses regarding main effects. If interaction effects are of interest to the researcher, this necessitates the employment of a factorial design. Although methodological solutions to the fixed-effect fallacy are certainly worth considering, it is clear that they can not be applied for all studies. As a result of these problems, it may be necessary to

resort to post hoc statistical manipulations for many studies in which a dependent variable is considered random.

Statistical Manipulations

Model simplification (pooling). As is mentioned above, there are two types of statistical manipulations that are generally accepted for use when the appropriate error term to test for the effect of interest is not directly available. These are model simplification and quasi F-ratios. Model simplification becomes an option when one can determine with a reasonable degree of confidence that an "excess" variance component contributes nothing to the EMS for the effect of interest. In this case, it is permissible to remove the variance component that is assumed to be zero from the EMS equation for the effect of interest. Once this is done, the researcher can choose an appropriate error term for the adjusted EMS and go on to perform a simple F-test on the data (Green & Tukey, 1960; Hopkins, 1983; Malgady et al., 1979; Winer, 1971).

Suppose that one chose to test the treatment effect for the P:T x I design presented in Table 5. In this case, one would have to provide support for either the hypothesis that $\sigma^2_{ti} = 0$ or that $\sigma^2_{p:t} = 0$ in order to simplify the model such that there is an appropriate error term with which to conduct an F-test on the treatment effect. The first line of support for the simplification of the model is logical; does past experience and knowledge of the properties of the variables in question lead the researcher to believe that the variance components in question should be zero?

Clearly, it is unlikely that there is no variance among the persons within the treatment groups, so the most reasonable term to attempt to eliminate from the model is σ^2_{ti}.

The hypotheses that $\sigma^2_{ti} = 0$ can be tested statistically by demonstrating that $MS_{ti}/MS_{pi:t}$ is nonsignificant at a liberal level of significance, such as $\alpha = .20$ or .30 (see Winer, 1971, p.379). The smaller the F-ratio, the more confidence one can have that the variance being tested is equal to zero. The focus on estimation of variance components provided by generalizability theory can provide additional information by allowing one to explicitly examine the estimated variance of an effect that is to be discarded for the purpose of model simplification.

Continuing with the example using the mean squares presented in Table 5, if one finds that the hypothesis that $\sigma^2_{ti} = 0$ is tenable, then the appropriate error term to test the null hypothesis for the treatment effect is $MS_{p:t}$. In general, the consensus is that model simplification is preferable to hypothesis testing with quasi F-ratios when conditions allow it (Hopkins, 1983; Malgady et al., 1979; Winer, 1971). In fact, Winer even recommends model simplification as a means of increasing power for cases in which there is an appropriate error term available but it has few degrees of freedom (less than 20). This increase in power is a result of the fact that the degrees of freedom associated with the interaction effects dropped from the model are included in the degrees of freedom for the residual term. Of course, the negative aspect of this is that any variance that is

associated with the interaction in actuality is also included in the residual term. Regardless of its pros and cons, if the data do not support model simplification, the researcher will have no recourse but to resort to some variant on a quasi F-ratio to test the hypothesis of interest.

Quasi F-ratios. Although it is not a precise statistical test, a quasi F-ratio, F', is generally considered to be a reasonable approximation to a true F-ratio (see Green & Tukey, 1960, p. 151; Satterthwaite, 1946). When there is no appropriate error term available to test a hypothesis, one can synthesize a suitable error term from a combination of EMS equations. Going back to the P:T x I example (Table 5), one can construct an appropriate error term to test the null hypothesis for the treatment effect by taking $MS_{ti} + MS_{p:t} - MS_{ip:t}$ (for general rules of construction of quasi F-ratios, see Winer, 1971, 375-378). For this design, the degrees of freedom for the denominator would be

$$\frac{(MS_{ti} + MS_{p:t} - MS_{pi:t})^2}{\frac{MS_{ti}^2}{df_{ti}} + \frac{MS_{p:t}^2}{df_{p:t}} + \frac{MS_{pi:t}^2}{df_{pi:t}}} .$$

Malgady et al. (1979) enumerate a number of problems with the use of quasi F-ratios. These problems include: (a) they only approximate the F distribution, (b) F' has less statistical power than F, (c) construction of the quasi F-ratio involves addition or subtraction of mean squares, which does not result in a true mean

square, (d) subtraction of mean squares can lead to negative variance estimates, and (e) the design may reduce to the analysis of binomial data, consequently violating the normality assumption of the ANOVA. Cochran (1951) and Clark (1973) have presented methods by which the latter two problems respectively can be circumvented. However, these methods have problems of their own that limit their utility (Malgady et al., 1979; Winer, 1971).

Milliken and Johnson (1984) point out that with unbalanced data, quasi F-ratios are imprecise not only because the degrees of freedom are approximated, but also because the mean squares making up the error term are not necessarily independently distributed.

Synopsis

Generalizability Theory

The perspective offered by generalizability theory can be very useful for clarifying the functions going on beneath the surface of the ANOVA. This perspective can be particularly useful for situations such as those presented in this thesis in which constraints on design (e.g., inability to completely nest variables where warranted) and the definition of dependent variables as random preclude the use of traditional F-statistics for analysis. Generalizability theory is not mathematical wizardry, but it forces the close examination of experimental design from start to finish and also provides valuable information about the performance of the dependent variable.

A drawback of generalizability theory is that, because it relies heavily on the estimation of variance components, a generalizability analysis is only as good as the estimates that it is based upon. Unfortunately, estimates of variance can be unstable unless samples are quite large. Furthermore, because persons are treated as a factor in a generalizability analysis, design matrices tend to have a very large number of columns. The large design matrices produced in many generalizability studies can make the use of some computer analysis programs impractical, such as those that perform maximum likelihood estimation (Bell, 1985; Chastain & Willson, 1986). Of course, the practicality of a given method of analysis is largely dependent upon the computing resources available to the researcher.

Unbalanced Designs

For the analyses performed in this thesis, computing difficulties were compounded further because of the unbalanced nature of the data. With unbalanced data, mean squares are no longer orthogonal and therefore they lose the property of additivity. Several methods are available to handle unbalanced designs and these have been described along with some of their characteristics. Each of these methods has its strong points and its drawbacks; none of them stand out clearly as being the "best" choice for use in an unbalanced generalizability study. The priorities of the researcher should guide the choice of an appropriate method.

The Fixed-Effect Fallacy

The fixed-effect fallacy is committed when one treats a factor in an ANOVA as fixed when it should logically considered random in order to make a desired inference. This fallacy is quite common in educational and psychological research and its consequences can vary from being an aesthetically displeasing departure from strict statistical logic to raising serious doubts about the validity of a study. It is very convenient statistically to treat dependent variables as fixed when calculating F-ratios; this is the default situation for most computerized statistical packages, such as $SPSS_X$. However, if generalization beyond the specific conditions of the dependent variable is intended, researchers must pay close attention to statistical considerations and conduct their studies such that this generalization is appropriate.

In this thesis, which utilizes the perspective of generalizability theory, methods of avoiding the fixed-effect fallacy in the statistical analysis of complex designs will be demonstrated. The designs to be analyzed are extensions of the P:T x I design described by Hopkins (1984) to the case in which there is nesting on the dependent variable.

CHAPTER III

METHODOLOGY

This chapter presents the methodology used to demonstrate an extension of the type of analysis described by Hopkins (1984) in which the dependent variable is considered to be random. The data to be analyzed in this study consist of the results of the June, 1989 British Columbia Algebra 12 school leaving examinations. Given the nature of these data, it is impossible to alter the basic nature of the design, only the complexity resulting from the number of nesting variables incorporated. Therefore, the focus of this chapter will be on the post hoc statistical manipulations that would be necessary in order to avoid the fixed-effect fallacy. To maintain continuity with Hopkins (1984), the first analysis conducted on the data will be similar to the one conducted by Hopkins (1984) for the design summarized in Chapter II, Table 5 (see p. 46) in which persons (P) nested within treatment groups (T) were crossed with test items (I), P:T x I. Subsequent designs will involve further nesting on the dependent variable.

Universe of Admissible Observations

In the designs utilized in this study, five variables were chosen with which to organize the sampled data for use in the analyses. These variables were:

(a) persons (P), consisting of the population of students eligible to write the June, 1989 British Columbia Algebra 12 exams. A sample of 267 students belonging to this population was

included in this study; of these students, 137 were male and 130 were female;

(b) gender (G), a nesting variable for the persons variable with two possible levels, male and female;

(c) items (I), consisting of the population of multiple choice items that could potentially be chosen to appear in the June, 1989 B.C. Algebra 12 exam. A sample of 55 such items were included in the study; these items were scored dichotomously;

(d) levels of cognitive complexity based on Bloom's (1956) *Taxonomy of Educational Objectives* (L) - this is a dichotomous nesting variable for the items facet with zero representing knowledge level items and one representing all items measuring higher order cognitive processes as classified by the British Columbia Ministry of Education;

(e) content areas (C), a nesting variable for the items facet with five levels: (1) trigonometry, (2) conics and quadratic systems, (3) exponents and logarithms, (4) polynomial functions, and (5) sequences/series and binomial expansions. One content area of the examination, complex numbers, was not included in the analysis because there were only three items within the area, all measuring higher order cognitive processes.

Universe of Generalization

Strictly speaking, arguments could be made against the definition of any of the facets in this study as random. If one accepted such arguments, then the universe of generalization would be limited to the specific samples of conditions upon which

the data were collected for this study. For example, the admissible population for the person factor is the population of students eligible to write the June, 1989 Algebra 12 examinations in British Columbia. The sample of students in this study was not randomly sampled per se, but was a sample of convenience drawn from 10 schools. However, Rogers and Bateson (1991) indicated that the "students attending these schools possessed diverse socio-economic backgrounds, reflect the gender distribution and major ethnic groups in the province, and represent the full range of achievement" (p. 169). This suggests that the sample could be considered a random sample from the population.

An argument can also be made against Loevinger's (1965) point that items on tests are not really randomly selected. Although items are sampled purposively, other items that are equally acceptable could be exchanged for the observed conditions. In fact, three sets of parallel items are devised for different administrations of the Algebra 12 examinations each year. In addition, the specifications for these examinations have remained essentially the same since their introduction in 1983. Thus, although any sample of items in a particular form of the Algebra 12 examination is not truly random, other samples of items are as acceptable as, or exchangeable with, those used. Based on de Finetti's (1964) concept of exchangeability, the sample of items can therefore be considered random.

The nesting facets are more clearly of a fixed nature. Consider the facet for content areas of algebra. The definition of

content areas within algebra is arbitrary and in an educational context, it is much more reasonable to sample individual items from within the content areas taught rather than to choose samples from the content areas themselves. A similar argument can be made for the facet representing the educational objectives measured by a test item. Of course, gender is clearly a fixed facet.

In summary, items (I) and persons (P) are random factors in this study. Gender (G), content domain (C), and level of cognitive complexity (L) are fixed. The nesting structure of the items within the blocking variables, content domain and level of cognitive complexity, for the Algebra 12 examination is presented in Table 7. Four designs were analyzed in the following order: (a) P:G x I, corresponding to Hopkins (1984); and three "extensions" (b) P:G x I:C, (c) P:G x I:L, and (d) P:G x I:CL.

Naturally, the least unbalanced of the four designs is P:G x I. The lack of balance is solely on the gender factor. Inspection of the marginal totals of Table 7 reveals that the unbalancing due to the introduction of the level-of-cognitive-complexity facet (L) is much smaller than that due to the introduction of the content domain facet (C). The greatest amount of imbalance occurs when both L and C are included in the design. As the main purpose of this study is demonstration, the progression will be made from the most simple model to the most complex model before any attempts at model simplification will be made.

Table 7

Nesting Structure of the Item Sample From the June, 1989 British Columbia Algebra 12 Examination

Content Area

Cognitive Level	Conics	Exp.	Seq.	Trig.	Poly.	Total
Knowledge	7	3	4	7	5	26
Understanding	6	4	4	11	4	29
Total	13	7	8	18	9	n = 55
GENDER 1 137						
GENDER 1 130						

Note. Descriptions of the content areas abbreviated in the headings of this table are as follows:

1). Conics - Conics and Quadratic Systems
2). Exp. - Exponents and Logarithms
3). Seq. - Sequences, Series, and Binomial Expansions
4). Trig - Trigonometry
5). Poly. - Polynomial Functions.

Knowledge represents knowledge level items; Understanding represents all items of greater cognitive complexity than knowledge level items.

Data Analysis

The data for this study were analyzed on an Amdahl 5860 mainframe computer using the MTS and the IBM VM systems at the University of Alberta. The program used to conduct the bulk of the analysis was the LOGIC subprogram of the $SPSS_X$ userproc, UANOVA (Taerum, 1986). This program uses Henderson's (1953) Method III for calculating sums of squares; it also outputs F-ratios and quasi F-ratios.

The EMS equations produced by LOGIC are calculated using Hartley's (1967) method of synthesis. Although the EMS equations output by LOGIC are identical to those output by the SAS procedure GLM (T. Taerum, personal communication, September 25, 1991), these estimates "disagree with those given in many textbooks and with those given by another widely used statistical package, BMDP" (Samuels, Casella, & McCabe, 1991, p. 798). This disagreement arises because LOGIC and GLM do not set the constraint that the sum of the effects for the interactions between fixed and random factors equal zero (Samuels et al., 1991). Samuels et al. argue that for most interpretations of the mixed model this constraint is reasonable; in generalizability theory, this constraint arises as a direct result of the definitions of score effects (Brennan et al., 1980). Failing to impose this constraint results in the variance components associated with the interactions in question remaining in the EMS equations from which they would otherwise be dropped. The retention of these terms biases some of the variance estimates obtained in this

study. However, the relative computational efficiency of the LOGIC program (Taerum, 1989) in comparison to accessible alternatives was a compelling point in its favour given the large design matrices resulting from a generalizability study of this magnitude.

Supplementary analyses. Due to concerns over the extent of bias of the variance components resulting from their calculation based on inappropriate expected mean squares, variances for two of the simpler designs (P:G x I and P:G x I:C) were estimated using the maximum likelihood program in the SAS procedure VARCOMP (SAS Institute, 1988) for comparison with the original estimates.

Model Simplification and Quasi F-Ratios

Following completion of the preliminary analyses, the four models were simplified as discussed in Chapter Two with the goal of conducting F-tests for the gender effect. There are three reasons for the simplification of complex models such as those described above. Firstly, it can allow the formation of appropriate error terms for "experimental" effects of interest and thus the avoidance of quasi F-ratios. Secondly, the use of simpler models reduces the likelihood of negative variance estimates and their associated problems. Lastly, it can result in an increase in statistical power due to an increase in the degrees of freedom for the residual term, assuming the simplification is appropriate and SS_{res} does not increase greatly. This last benefit will not be attained in the analysis in question, however, because SS_{res} is not the error term for the gender effect (see p. 80), so changes in its

degrees of freedom will not affect power for the analysis of interest.

The use of quasi F-ratios provides an alternative to model simplification for testing the significance of effects for which there is no appropriate error term. Following the model simplification, quasi F-ratios were calculated for the four designs, allowing comparison of the results of each method of analysis.

The variance estimates resulting from the preliminary analysis of the four designs described above and their confidence intervals are presented in the following chapter. Inferential tests involving both the use of model simplification and quasi F-ratios to examine the significance of the gender effect are also illustrated. Finally, generalizability coefficients for the examination are calculated based on the obtained variance estimates.

CHAPTER IV

RESULTS

Item level results of the June, 1989 British Columbia Algebra 12 leaving examinations were the data used in the analyses of four experimental designs: P:G x I, P:G x I:C, P:G x I:L, and P:G x I:CL. These data were analyzed from the perspective of generalizability theory with persons (P) and items (I) treated as random factors, and gender (G), content domain (C), and level of cognitive complexity (L) treated as fixed.

Organization of the Chapter

The results of this study are presented in the sequence in which a generalizability study is conducted. First, the expected mean squares calculated using the method of synthesis (Hartley, 1967) are presented together with a description of how they differ from the expected mean squares that would be calculated based on common algorithms for the balanced mixed model (eg., Winer, 1971, pp. 371-375). These differences are brought about by (a) the failure to impose the constraint that the sum of score effects for interactions involving fixed effects be zero and (b) the nonorthogonality resulting from the unequal n's.

Following the presentation of the expected mean squares, the variance components estimated by equating the EMS with the obtained mean squares are presented along with their associated confidence intervals. These variance components are then combined to form generalizability coefficients for the four ways of conceptualizing the structure of the Algebra 12 examination

presented herein. Finally, inferential statistics are performed on the gender effect, first using the model simplification approach and then using quasi F-ratios.

Expected Mean Squares

The expected mean square equations for the four designs considered in this study that were calculated by LOGIC (Taerum, 1986) by the application of Hartley's (1967) synthesis procedure to the design matrices are presented in Tables 8 to 11. To illustrate the complexities introduced with an unbalanced design, additional terms to those expected in the balanced case that result from the nonorthogonality of the data set have been typeset in bold face. Variance components that appear in the EMS equations due to the failure to impose the constraints of the generalizability model are underlined in these tables. The remaining terms in these equations are those that would appear in a balanced mixed model for each design had the constraints of generalizability theory been imposed. For ease of reference, the EMSs for the four designs, calculated under the constraints of generalizability theory and assuming a balanced design, are presented in Appendix A.

As is shown in Table 8, the EMSs for the P:G x I design contain one variance component (underlined) in $E(MS_I)$ that should not have appeared had the constraints of generalizability theory been imposed. Although the result, $E(MS_I) = \sigma^2_{ip:g} + 133.588\sigma^2_{ig} + 267.000\sigma^2_{i}$ was obtained, the appropriate equation is $E(MS_I) = \sigma^2_{ip:g} + 267.000\sigma^2_{i}$ (based on Winer, 1971, pp. 371-375). As more fixed effects are incorporated into the designs, more inappropriate

terms are included in the models. Thus, when L or C are included as facets, four inappropriate terms appear (see Tables 9 and 10); when both L and C are included, there are twenty-five inappropriate terms included in the EMS equations (see Table 11). Although the total number of terms is greater in the more complex designs, the proportion of inappropriate terms and, thus, the threat of bias is clearly greater for such designs.

Table 8
Mean Square Model for the P:G x I Design

SOURCE	EMS
G	$\sigma^2_{ip:g} + 133.407\sigma^2_{ig} + 55.001\sigma^2_{p:g} + 7337.440\sigma^2_{g}$
P:G	$\sigma^2_{ip:g} + 55.000\sigma^2_{p:g}$
I	$\sigma^2_{ip:g} + 133.588\sigma^2_{ig} + 267.000\sigma^2_{i}$
IG	$\sigma^2_{ip:g} + 133.411\sigma^2_{ig}$
IP:G	$\sigma^2_{ip:g}$

Turning now to changes brought about by nonorthogonality, additional variance components were obtained in the EMSs for the P:G x L and P:G x I:CL designs. In the first instance (Table 9), the additional term was $0.001\sigma^2_{i:l}$ for $E(MS_L)$. In the second instance (Table 11), several additional variance components arise (eg., see

EMS for L, C, G x L, G x C, and PL:G). This suggests that the dependence is most pronounced in the most complex model as would be expected.

Inspection of the components brought about by nonorthogonality (bold face) reveals that each contains at least one of the nesting facets for items and each is included in an EMS equation involving one of these nesting facets. Furthermore, for all but one of these components [$0.001\sigma^2_{i:l}$ in $E(MS_L)$], if the source of variance includes the C facet, then the additional component includes L and vice versa. For example, $E(MS_{GL})$ contains additional terms containing $\sigma^2_{i:cl}$, $\sigma^2_{pc:g}$, and σ^2_{gc}. Comparing the additional components due to nonorthogonality (bold face) with those that would be included in a balanced design (plain type) reveals that they are relatively small.

Table 9
Mean Square Model for the P:G x I:C Design

SOURCE	EMS
G	$\sigma^2_{ip:gc} + \underline{12.489}\sigma^2_{pc:g} + \underline{1666.380}\sigma^2_{gc} + 133.407\sigma^2_{ig:c} + 55.001\sigma^2_{p:g} + 7337.440\sigma^2_{g}$
C	$\sigma^2_{ip:gc} + 10.626\sigma^2_{pc:g} + \underline{133.590}\sigma^2_{ig:c} + 266.999\sigma^2_{i:c} + \underline{1419.710}\sigma^2_{gc} + 2837.480\sigma^2_{c}$
P:G	$\sigma^2_{ip:gc} + \underline{12.491}\sigma^2_{pc:g} + 55.000\sigma^2_{p:g}$
I:C	$\sigma^2_{ip:gc} + \underline{133.589}\sigma^2_{ig:c} + 267.000\sigma^2_{i:c}$
GC	$\sigma^2_{ip:gc} + 10.627\sigma^2_{pc:g} + 133.409\sigma^2_{ig:c} + 1417.760\sigma^2_{gc}$
PC:G	$\sigma^2_{ip:gc} + 10.627\sigma^2_{pc:g}$
IG:C	$\sigma^2_{ip:gc} + 133.411\sigma^2_{ig:c}$
IP:GC	$\sigma^2_{ip:gc}$

Table 10
Mean Square Model for the P:G x I:L Design

SOURCE	EMS
G	$\sigma^2_{ip:gl} + \underline{27.581}\sigma^2_{pl:g} + 133.407\sigma^2_{ig:l} + \underline{3679.630}\sigma^2_{gl} + 55.001\sigma^2_{p:g} + 7337.440\sigma^2_{g}$
L	$\sigma^2_{ip:gl} + 27.417\sigma^2_{pl:g} + \underline{133.589}\sigma^2_{ig:l} + 266.997\sigma^2_{i:l} + 7320.640\sigma^2_{l}$
P:G	$\sigma^2_{ip:gl} + \underline{27.580}\sigma^2_{pl:g} + 55.000\sigma^2_{p:g}$
I:L	$\sigma^2_{ip:gl} + \underline{133.589}\sigma^2_{ig:c} + 267.000\sigma^2_{i:l}$
GL	$\sigma^2_{ip:gl} + 27.418\sigma^2_{pl:g} + 133.409\sigma^2_{ig:l} + 0.001\sigma^2_{i:l} + 3657.820\sigma^2_{gl}$
PL:G	$\sigma^2_{ip:gl} + 27.420\sigma^2_{pl:g}$
IG:L	$\sigma^2_{ip:gl} + 133.411\sigma^2_{ig:l}$
IP:GL	$\sigma^2_{ip:gl}$

Table 11

Mean Square Model for the P:G x I:CL Design

SOURCE	EMS
G	$\sigma^2_{ip:gcl} + \underline{6.418}\sigma^2_{pcl:g} + 133.407\sigma^2_{ig:cl} + \underline{856.238}\sigma^2_{gcl} + \underline{12.489}\sigma^2_{pc:g} + \underline{27.581}\sigma^2_{pl:g} + \underline{1666.380}\sigma^2_{gc} + 55.001\sigma^2_{p:g} + \underline{3679.630}\sigma^2_{gl} + 7337.440\sigma^2_{g}$
L	$\sigma^2_{ip:gcl} + \underline{133.589}\sigma^2_{ig:cl} + \underline{6.343}\sigma^2_{pcl:g} + 27.417\sigma^2_{pl:g} + 0.270\sigma^2_{i:c} + 847.383\sigma^2_{gcl} + 266.996\sigma^2_{i:cl} + \underline{3662.830}\sigma^2_{gl} + 36.115\sigma^2_{gc} + \underline{1693.600}\sigma^2_{cl} + 7320.640\sigma^2_{l}$
C	$\sigma^2_{ip:gcl} + \underline{133.590}\sigma^2_{ig:cl} + \underline{5.421}\sigma^2_{pcl:g} + 10.626\sigma^2_{pc:g} + 0.130\sigma^2_{gl} + \underline{724.458}\sigma^2_{gcl} + 266.999\sigma^2_{i:cl} + \underline{1447.910}\sigma^2_{cl} + 17.636\sigma^2_{gl} + \underline{1419.710}\sigma^2_{gc} + 2837.480\sigma^2_{c}$
~PG	$\sigma^2_{ip:gcl} + \underline{6.417}\sigma^2_{clp:g} + \underline{27.580}\sigma^2_{pl:g} + \underline{12.491}\sigma^2_{pc:g} + 55.000\sigma^2_{p:g}$
I:LC	$\sigma^2_{ip:gcl} + \underline{133.588}\sigma^2_{ig:c} + 267.000\sigma^2_{i:lc}$
GL	$\sigma^2_{ip:gcl} + \underline{6.338}\sigma^2_{pcl:g} + 133.408\sigma^2_{ig:cl} + \underline{846.223}\sigma^2_{gcl} + 0.002\sigma^2_{i:cl} + 0.268\sigma^2_{pc:g} + 27.418\sigma^2_{pl:g} + 36.071\sigma^2_{gc} + 3657.820\sigma^2_{gl}$

Table 11 (Continued)

SOURCE	EMS
GC	$\sigma^2_{ip:gcl} + \underline{5.423}\sigma^2_{pcl:g} + 133.409\sigma^2_{ig:cl} + \underline{723.459}\sigma^2_{gcl} + \mathbf{0.132}\sigma^2_{\mathbf{pl:g}} + 10.627\sigma^2_{pc:g} + 1417.760\sigma^2_{cg}$
LC	$\sigma^2_{ip:gcl} + \underline{133.590}\sigma^2_{ig:c} + 5.135\sigma^2_{pcl:g} + 267.001\sigma^2_{i:cl} + \underline{686.222}\sigma^2_{gcl} + 1371.510\sigma^2_{cl}$
LCG	$\sigma^2_{ip:gcl} + 5.138\sigma^2_{pcl:g} + 133.409\sigma^2_{ig:cl} + 685.295\sigma^2_{gcl}$
PLG	$\sigma^2_{ip:gcl} + \underline{6.343}\sigma^2_{pcl:g} + \mathbf{0.269}\sigma^2_{\mathbf{pc:g}} + 27.420\sigma^2_{pl:g}$
PCG	$\sigma^2_{ip:gcl} + \underline{5.421}\sigma^2_{pcl:g} + 10.627\sigma^2_{pc:g}$
IG:LC	$\sigma^2_{ip:gcl} + 133.412\sigma^2_{ig:cl}$
CLP:G	$\sigma^2_{ip:gcl} + 5.139\sigma^2_{clp:g}$
IP:GCL	$\sigma^2_{ip:gl}$

Estimates of Variance Components

In a generalizability study, the results of primary interest are the variance estimates for the random components of the model. These were calculated by equating the expected mean squares reported in Tables 8 to 11 with the computed mean

squares obtained from the ANOVA. The variance estimates obtained are presented in Table 12 for each of the four designs.

Inspection of Table 12 reveals that the variance components common to the four designs are essentially stable, with the exception of variance due to items. When the level of cognitive complexity (L) is included in the design (Designs 2 & 4), $\hat{\sigma}^2_i$ becomes smaller. Comparison of the mean scores for the items classified at the knowledge level and for the items classified as requiring greater cognitive complexity reveals that the former is substantially greater than the latter (.660 compared to .461). Such a difference in mean scores among content domains was not observed. From a learning perspective, solution of items of greater cognitive complexity requires a sound knowledge base in their content area; thus if higher level items are answered correctly, it is probable that knowledge level items will be answered correctly as well, but the reverse is not necessarily true. Content domains, on the other hand, do not exhibit this hierarchical nature and thus it is more likely that their means will be similar to each other.

Looking now at the decomposition of the total variance, the largest variance components across the four models are attributable to (a) the item factor , in which, as pointed out, a sizable portion of the variance can be attributed to differences in cognitive complexity among the items, (b) the person factor, and (c) the item by person interaction. Most of the remaining terms have negligible amounts of variance associated with them. A

negative variance estimate was computed for the term $\hat{\sigma}^2_{pI:g}$ in the P:G x I:CL design. This estimate was retained for the computation of variance estimates for terms in which $\sigma^2_{pI:g}$ contributes to the EMS.

Assessment of Possible Bias

As was discussed previously, some of the EMSs utilized in this study for the estimation of variance components contain terms that are inappropriate for a generalizability analysis of a mixed model. The presence of these inappropriate terms leads to biased estimates of variance, although this bias is likely to be small due to the small size of the variances associated with these terms, the largest being $\hat{\sigma}^2_{pc:g} = 0.003$. Assuming non-negative variance estimates, the bias will be in a downward direction for the highest level interaction containing an inappropriate term, but can be in any direction for higher terms containing a biased variance component in their EMS. For the particular designs in this study, however, all of the bias is in a downward direction. As an example of the magnitude of this bias, $\hat{\sigma}^2_{p:g}$ was calculated for the P:G x I:C design both with the inappropriate term, $12.491\sigma^2_{pc:g}$ retained and with it removed (see Appendix B). This component was chosen as an example because it is biased by $\hat{\sigma}^2_{pc:g}$, so it should provide a fair indication of the magnitude of bias. The unbiased estimate obtained was 0.01327 in comparison with the biased estimate of 0.01263, a negligible difference. Biased estimates are flagged in Table 12 with asterisks.

Table 12

Variance Estimates Resulting From the Analysis of Four Designs Using Henderson's Method III

	P:G x I	P:G x I:C	P:G x I:L	P:G x I:CL
$\hat{\sigma}^2_i$	0.048*	0.049*	0.039*	0.039*
$\hat{\sigma}^2_{p:g}$	0.015	0.013*	0.015*	0.013*
$\hat{\sigma}^2_{ig}$	2.95×10^{-4}	2.72×10^{-4}	3.13×10^{-4}	2.15×10^{-4}
$\hat{\sigma}^2_{ip:g}$	0.184	0.182	0.184	0.181
$\hat{\sigma}^2_{pc:g}$	N/A	0.003	N/A	0.003*
$\hat{\sigma}^2_{pl:g}$	N/A	N/A	1.47×10^{-4}	-2.00×10^{-4}*
$\hat{\sigma}^2_{clp:g}$	N/A	N/A	N/A	8.04×10^{-4}

Note. It is assumed in the above table that wherever the item facet is indicated, it is nested within the facets appropriate to the given design. Thus, for:
1) the P:G x I:C design, $\hat{\sigma}^2_i$ represents $\hat{\sigma}^2_{i:c}$, $\hat{\sigma}^2_{gi}$ represents $\hat{\sigma}^2_{gi:c}$, and $\hat{\sigma}^2_{ip:g}$ represents $\hat{\sigma}^2_{ip:gc}$,
2) the P:G x I:L design, $\hat{\sigma}^2_i$ represents $\hat{\sigma}^2_{i:l}$, $\hat{\sigma}^2_{gi}$ represents $\hat{\sigma}^2_{gi:l}$, and $\hat{\sigma}^2_{ip:g}$ represents $\hat{\sigma}^2_{ip:gl}$, and
3) the P:G x I:CL design, $\hat{\sigma}^2_i$ represents $\hat{\sigma}^2_{i:cl}$, $\hat{\sigma}^2_{gi}$ represents $\hat{\sigma}^2_{gi:cl}$, and $\hat{\sigma}^2_{ip:g}$ represents $\hat{\sigma}^2_{ip:gcl}$.
N/A indicates that the given variance component is not applicable to the design in question.
* denotes biased estimates.

To further assess the magnitude of the bias resulting from the use of inappropriate variance components within the EMS equations used to estimate the variance components in this study, supplementary variance estimates were obtained using maximum likelihood (ML) estimation for two designs: P:G x I and P:G x I:C.

Although the full data set was used in the analysis of the first design, the sample size for persons was reduced to 20 males and 16 females for the ML analysis of the P:G x I:C design due to computer memory limitations. To have a basis with which to compare the performance of LOGIC and ML for the P:G x I:C design, LOGIC was employed to obtain variance estimates on this reduced data set also. The variance components obtained from these analyses are presented in Tables 13 and 14.

The comparison between the results of the two analyses in each pair is complicated by the fact that ML estimates are also biased. This bias is generally in a downward direction, but is also subject to upward pressure due to the limitation on the parameter space for variance to non-negative values. This upward pressure minimizes bias for variance estimates that are small relative to the error variance (Swallow & Monahan, 1984).

For the P:G x I design (Table 13) the results of the Method III and ML analyses agree very closely, with the ML estimates tending to be very slightly lower than the Method III estimates. This provides further evidence that for the P:G x I design the addition of the inappropriate variance component in the EMS equation for items (Table 8) does not significantly affect the variance estimates. For the P:G x I:C design (Table 14), the estimates from the two methods diverge slightly for some estimates. For two estimates, $\hat{\sigma}^2_{gi:c}$ and $\hat{\sigma}^2_{pc:g}$, there was more than a 30% difference between the results from ML and Method III, although the difference was still only 0.001 in each case. The

insignificance of these differences becomes clear when one realizes that for both of these components, the actual variances may be zero.

Given that one would expect to find differences due to different methods of estimation, the differences found in Tables 13 and 14 are small. However, the bias in the Method III results may become slightly more noticeable for the most complex analysis due to the greater number of inappropriate terms included in the EMS equations used to arrive at the variance estimates. Regardless of this, the similarity of variance estimates across designs in Table 11 implies that the bias resulting from the use of Method III is of a small magnitude, even for the most complex design in this study.

Table 13

Comparison of Variance Estimates Derived From Method III and ML Analyses for the P:G x I Design

	P:G x I (Method III)	P:G x I (ML)
$\hat{\sigma}^2_i$	0.048*	0.047
$\hat{\sigma}^2_{p:g}$	0.015	0.015
$\hat{\sigma}^2_{gi}$	2.95×10^{-4}	2.85×10^{-4}
$\hat{\sigma}^2_{ip:g}$	0.184	o.183

Table 14

Comparison of Variance Estimates Derived From Method III and ML Analyses for the P:G x I:C Design on a Reduced Data Set

	P:G x I:C (Method III)	P:G x I:C (ML)
$\hat{\sigma}^2_{i:c}$	0.045	0.042
$\hat{\sigma}^2_{p:g}$	0.011	0.011
$\hat{\sigma}^2_{gi:c}$	0.004	0.003
$\hat{\sigma}^2_{ip:gc}$	0.180	0.181
$\hat{\sigma}^2_{pc:g}$	0.005	0.004

Stability of Variance Estimates

The stability of variance estimates can be assessed by examination of confidence intervals constructed around those estimates. Approximate confidence intervals around the variance estimates in Table 12 are provided in Table 15. These intervals are calculated using Satterthwaite's (1946) procedure, which is presented in Appendix C. In the most complex design, P:G x I:TC, the confidence interval noticeably widens relative to those in the simpler designs only for the gender by item interaction. The estimate $\hat{\sigma}^2_{clp:g}$ also has a wide confidence interval in this design, but there is no analog in the simpler designs with which to

Table 15

Approximate 90% Confidence Intervals for the Variance Estimates Resulting From the Application of Four Models

P:G x I

	df	CI
$\hat{\sigma}^2_{i}$	52.15	0.04 - 0.07
$\hat{\sigma}^2_{p:g}$	178.01	0.01 - 0.02
$\hat{\sigma}^2_{gi}$	1.67	0.00 - 0.01
$\hat{\sigma}^2_{ip:g}$	14310	0.18 - 0.19

P:G x I:C

	df	CI
$\hat{\sigma}^2_{i:c}$	48.36	0.04 - 0.07
$\hat{\sigma}^2_{p:g}$	150.20	0.01 - 0.02
$\hat{\sigma}^2_{gi:c}$	1.39	0.00 - 0.02
$\hat{\sigma}^2_{ip:gc}$	13250	0.18 - 0.19
$\hat{\sigma}^2_{pc:g}$	21.85	0.00 - 0.01

P:G x I:L

	df	CI
$\hat{\sigma}^2_{i:l}$	50.74	0.03 - 0.06
$\hat{\sigma}^2_{p:g}$	170.48	0.01 - 0.02
$\hat{\sigma}^2_{gi:l}$	1.81	0.00 - 0.01
$\hat{\sigma}^2_{ip:gl}$	14045	0.18 -0.19
$\hat{\sigma}^2_{pl:g}$	0.12	0.00 - 0.00

P:G x I:CL

	df	CI
$\hat{\sigma}^2_{i:cl}$	43.22	0.03 - 0.06
$\hat{\sigma}^2_{p:g}$	144.98	0.01 - 0.02
$\hat{\sigma}^2_{gi:cl}$	0.48	0.00 - 0.15
$\hat{\sigma}^2_{ip:gcl}$	11925	0.18 - 0.19
$\hat{\sigma}^2_{pc:g}$	10.49	0.00 - 0.01
$\hat{\sigma}^2_{pl:g}$	0.18	0.00 - 0.00
$\hat{\sigma}^2_{pcl:g}$	0.49	0.00 - 65.55

compare it. These wide confidence intervals can be considered artifacts of the very small "effective" degrees of freedom for these components, being 0.48 and 0.49 respectively.

Inferential Statistics

Model Simplification

The goal of the model simplification presented here is to alter the model such that the significance of the gender effect can be assessed with a simple F-test. As the EMSs presented in Tables 8 to 11 are based on assumptions that are not appropriate for testing main effects (Samuels et al., 1991), the EMS tables relevant to this discussion are presented in Appendix A (Tables A to D). Examination of the EMSs presented in these tables reveals that for each of the four designs, there is no existing error term with which to to test the gender effect. However, in all of these designs, if the variance component due to the item by gender interaction is removed from the EMS for gender, then EMS(P:G) becomes an appropriate error term.

Logically, one would expect (and hope) that a provincial examination that is used to rank students academically does not suffer from gender related item bias, or in other words that $\sigma^2_{gi} = 0$. Summary ANOVA tables for the four designs are presented in Tables 16 to 19. Note that in all of these cases, the gender by item interaction is nonsignificant.

Table 16

ANOVA Table for the Analysis of the P:G x I Design

SOURCE	SS	DF	MS	F	p
G	10.17	1	10.17	-	
I	703.60	54	13.03	72.39	< .05
P:G	270.49	265	1.02	5.67	< .05
GI	12.05	54	0.22	1.22	0.14
IP:G	2631.54	14310	0.18	-	

Table 17

ANOVA Table for the Analysis of the P:G x I:C Design

SOURCE	SS	DF	MS	F	P
G	7.26	1	7.26	-	
C	39.47	4	9.87	-	
GC	1.15	4	0.29	-	
I:C	664.12	50	13.28	60.96	< .05
P:G	241.25	265	0.91	4.27	< .05
PC:G	225.81	1060	0.21	1.17	< .05
GI:C	10.90	50	0.22	1.20	0.15
IP:GC	2405.73	13250	0.18	-	

Table 18

ANOVA Table for the Analysis of the P:G x I:L Design

SOURCE	SS	DF	MS	F	P
G	10.25	1	10.25	-	
L	143.82	1	143.82	-	
GL	0.09	1	0.09	-	
I:L	559.78	53	10.56	46.82	< .05
P:G	270.17	265	1.02	5.43	< .05
PL:G	49.78	265	0.19	1.02	0.39
GI:L	11.96	53	0.23	1.23	0.12
IP:GL	2581.76	14045	0.18		

Although this effect is nonsignificant at $\alpha = .05$, the obtained p values, with the exception of that for GI:CL tend to be lower than the .20 to .30 suggested by Winer (1971) as a minimum value for dropping a component from the model. Winer's suggested α levels are general guidelines, however. Green and Tukey (1960) favored using the magnitude of an effect as a criterion for deciding on whether or not pooling is acceptable, suggesting that pooling is appropriate when the F-ratio of the given term is less than two (p. 138). With 14,310 degrees of freedom in the denominator of the F-ratio in question, the effect size becomes a consideration of primary importance. As one would expect from the small F-ratio obtained for the gender by item interaction, $\hat{\sigma}^2_{gi}$ is near zero and

Table 19

ANOVA Table for the Analysis of the P:G x I:CL Design

SOURCE	SS	DF	MS	F	P
G	7.10	1	7.10	-	
C	36.82	4	9.20	-	
L	119.24	1	119.24	-	
GC	1.09	4	0.27	-	
GL	0.07	1	0.07	-	
CL	47.22	4	11.80	-	
GCL	1.32	4	0.33	-	
I:CL	476.99	45	10.60	50.54	< .05
P:G	243.06	265	0.92	5.06	< .05
PC:G	227.16	1060	0.21	1.17	0.06
PL:G	48.08	265	0.18	1.00	0.50
PCL:G	196.34	1060	0.19	1.02	0.30
GI:CL	9.44	45	0.21	1.16	0.22
IP:GCL	2159.50	11925	0.18		

very small relative to the other components in the model. Given this, dropping this interaction from the model seems reasonable.

F-ratios for the gender effect based on the results of the Method III analysis for the four designs are presented in Table 20.

Due to the nature of the Method III analysis, or fitting constants method, mean squares are independent of each other, allowing valid construction of F-ratios.

Table 20

F-ratios for the Gender Effect in Four Experimental Designs

DESIGN	MSb	MSw	df	F
P:G x I	10.17	1.02	1, 265	9.97
P:G x I:C	7.26	0.91	1, 265	7.98
P:G x I:L	10.25	1.07	1, 265	10.04
P:G x I:CL	7.10	0.95	1, 265	7.71

Quasi F-ratios

If one decides that it is unacceptable to drop the item by gender interaction term from the model, then the only alternative remaining for testing significance of the gender effect is the use of quasi F-ratios. Quasi F-ratios calculated based on the Method III results for the four designs according to the method discussed in Chapter II are presented in Table 21. Relative to the F-ratios presented in Table 20, the quasi F-ratios are slightly smaller in magnitude, due to their slightly larger error mean squares. They also have fewer degrees of freedom for the error term in all four designs, resulting in decreased statistical power.

Table 21

Quasi F-ratios for the Gender Effect in Four Experimental Designs

	MSo	MSw	df	F'
P:G x I	10.17	1.06	1, 233.0	9.59
P:G x I:C	7.26	0.95	1, 220.4	7.64
P:G x I:L	10.25	1.07	1, 232.4	9.58
P:G x I:CL	7.10	0.95	1, 214.8	7.47

Generalizability Coefficients

One advantage of using generalizability theory in the analysis of data collected in a study is that generalizability coefficients for given applications of the dependent measure can be estimated from the G-study data. The primary goal of the B.C. Algebra 12 examinations is the ranking of students, so these examinations can be considered norm referenced. Therefore, the appropriate type of generalizability coefficient to calculate from the data is one with a relative error term. As described in Chapter II, the general formula for this type of generalizability coefficient is $\hat{\rho} = \hat{\sigma}^2_u/[\hat{\sigma}^2_u + \hat{\sigma}^2_\delta]$, where $\hat{\sigma}^2_u$ represents universe score variance and $\hat{\sigma}^2_\delta$ represents relative error variance.

For all of the designs presented in this thesis, $\hat{\sigma}^2_{pi}$ is the only term contributing to relative error variance. The terms

contributing to universe score variance differ among the designs, however. This is because the interaction of persons with fixed facets contributes to universe score variance. The universe score variances associated with each of the four designs analyzed in this thesis are presented below:

1) for P:G x I, $\hat{\sigma}^2_u = \hat{\sigma}^2_p$,

2) for P:G x I:C, $\hat{\sigma}^2_u = \hat{\sigma}^2_p + \hat{\sigma}^2_{pc}/n_c$,

3) for P:G x I:L, $\hat{\sigma}^2_u = \hat{\sigma}^2_p + \hat{\sigma}^2_{pl}/n_l$,

4) for P:G x I:CL, $\hat{\sigma}^2_u = \hat{\sigma}^2_p + \hat{\sigma}^2_{pc}/n_c + \hat{\sigma}^2_{pl}/n_l + \hat{\sigma}^2_{pcl}/n_c n_l$.

The generalizability coefficients computed from the four designs are presented in Table 22. All of these coefficients differ from each other by less than 0.02, regardless of differences in design used for the estimation of variance components. The magnitude of these deviations due to differences in design are similar to those obtained by Chastain & Willson (1986) in their analysis of WAIS-R data.

The generalizability coefficients obtained in this study are somewhat smaller than the reliability coefficient of 0.84 obtained from a previous study (Rogers & Bateson, 1991, p. 171). Part of the reason for this is that only 55 of 58 items were used in this study, but even after the item sample size is adjusted using the Spearman-Brown prophecy formula, a reliability coefficient of 0.829 is obtained for the P:G x I design. The remaining difference

Table 22

Estimates of Generalizability Coefficient for Four Designs

	$\hat{\rho}$
P:G x I	.817
P:G x I:C	.804
P:G x I:L	.818
P:G x I:CL	.805

may be attributed to the effect of breaking up the sample of students into males and females. Males and females can be considered intact groups that differ in central tendency; the previously obtained reliability coefficient of .84 may have been inflated due to the pooling of these groups.

CHAPTER V

DISCUSSION

In this thesis, the work of Hopkins (1984) in which he integrates generalizability theory with ANOVA for an experiment-like design was extended to include items of a test as a random sample stratified by content domain and level of cognitive complexity (ie., as a nested, random facet). The perspective of generalizability theory was taken for two reasons. Firstly, it requires one to explicitly define the universe of admissible observations and the universe of generalization; this approach makes it clear that if one wishes to generalize beyond the particular measurement instrument used, then one must treat the dependent variable as a random factor. Secondly, incorporating generalizability theory into the data analysis allows one to examine the performance of the dependent measure on the population of interest by focusing attention on the variance components associated with the facets that contribute to measurement error in addition to the performance of inferential statistics.

Although generalizability theory is very useful for the reasons mentioned above, there are also problems associated with it. The introduction of the dependent measure as a random facet, although logically required for valid inferences beyond the particular measuring device used, makes it unlikely that a suitable term will exist with which to test the effects of interest. Also, the examination of the dependent measure encouraged in

generalizability theory will often reveal that the item facet is stratified such that the design is unbalanced, resulting in nonorthogonality. Furthermore, like any sample-based estimates, the variance estimates that are central to generalizability theory are not stable. Thus, relatively large sample sizes are required to obtain usable results, especially with a complex design.

There are also difficulties associated with the computing resources required to conduct generalizability analyses. The treatment of persons as a factor, a central feature of the generalizability analysis, results in large design matrices, thereby increasing the memory and c.p.u. time required to perform an analysis relative to a traditional ANOVA. These large design matrices can cause the requirements in computer resources to become prohibitive when methods of analysis requiring matrix inversion, such as maximum likelihood analysis, are used. Furthermore, most ANOVA programs do not produce variance estimates, so unless it is feasible to calculate the estimates from the mean squares, the choice of packages with which to analyze the data is limited.

Due to difficulties associated with nonorthogonality, decreased accuracy in the estimation of variance components, and the requirement for extensive computing resources, the analysis of unbalanced data should be avoided if possible, especially within the context of generalizability theory. In many cases, the option of randomly discarding data to achieve a balanced design should be given serious consideration (Brennan, 1983). In this study, for

instance, it would have been quite reasonable to discard seven of the male students to artificially balance the P:G x I design. This option is not always feasible, however. For instance, in the most complex design, P:G x I:CL, 25 of 55 items would have to be discarded, leaving three items per cell. Given the large item variance, this would not be acceptable.

Significance Tests for Differences Among Blocking Factors on the Dependent Variable

The data that Coleman (1973) reanalyzed to find that "no reliable evidence for most of the main conclusions drawn from them" (p. 335) were structured such that samples of words were nested within different classes (eg., homophones versus nonhomophones). A similar analysis to the above for the Algebra 12 data presented in this thesis would be the analysis of whether there are mean differences in p values for items nested within different content domains or within different levels of Bloom's (1956) taxonomy. Clearly, we would be well advised to try to avoid the possibility that differences in mean scores for different conditions of the nesting variables are due to the particular items chosen.

A general principle can be inducted from these examples. If a goal of a study is the comparison of the effects of different levels of a blocking factor for the dependent variable, then the fixed-effect fallacy is a serious threat to external validity and this threat must be explicitly addressed.

To test hypotheses such as those presented above, the procedures suggested by Hopkins (1984) and incorporated into this thesis should be utilized. Although it is very unlikely that the variance of items within blocking factors would be zero, the interaction of persons with the blocking factors could quite possibly be eliminated, leaving MS_I as the error term. In this case, it is crucial for the sake of power that several levels of the dependent measure be randomly sampled within the nesting variable because the analysis of any differences among these two levels is the crux of the study. The greater the variance within the blocks, the greater the sample size needed for the nested item variable. In such a situation, it would be valuable to first conduct a G-study to estimate the variance within the blocks and thus illuminate the decision as to the size of the item sample needed. Note that for this case, persons must be defined as a factor to permit estimation of an interaction term that enables simplification to an F-ratio or construction of a quasi F-ratio.

Significance Tests for Differences Among Conditions of the Independent Variable(s)

A common case for studies in the educational and psychological literature is one in which the group means on a test or inventory are compared for the different treatment groups, as was done by Hopkins (1984). An example of this type of case would be an analysis of the treatment effect, gender (G) in the design P:G x I. In this case, it is the interaction between the

treatment effect and the dependent variable, otherwise known as item bias, that is the threat to external validity.

Regardless of how the data are analyzed, it is clearly undesirable to have a measure that is biased towards a given group in such a design and ideally, the researcher should do whatever can be done to attempt to eliminate the bias before proceeding with the study rather than attempting to accommodate it statistically through the use of quasi F-ratios. Although most well standardized measures have been developed such that bias is avoided and σ^2_{gi} can be logically assumed to be zero for the universe of persons on which their use is intended, testing this assumption statistically before simplifying to an essentially fixed-effects model can add some strength to the case for such a model.

Complexity of Experimental Designs

Smith (1978) demonstrated that with increasing complexity of design, estimates of variance components become less accurate. In this study, however, it was found that increasing the levels of nesting on the dependent variable had a minimal effect on the confidence intervals around the variance estimates for the random effects. There are two main reasons behind this apparent discrepancy in results. Firstly, all of the nesting facets in the analyses presented herein are fixed. Had they been random, the large number of variance terms included in their expected mean squares would have made them quite susceptible to dispersion effects, which cause instability. Secondly, all of the interaction

components that were included in the EMSs for the terms for which variances were estimated were essentially zero (other than the trivial case of the residual variance), so dispersion due to errors in the estimation of these components was minimal.

In addition, all four of the designs analyzed had essentially two crossed factors, persons and items, with the increase in complexity coming about as a result of nesting on the second of these factors. Although this has little relevance to the stability of the estimates obtained in this study, it would be highly relevant in a more ideal case. Had the variances been estimated from EMS equations appropriate to mixed model generalizability analyses, the EMS equations associated with the effects estimated would have all been very simple (see Appendix A), resulting in minimal dispersion effects and a minimal reduction is stability with the addition of more fixed blocking factors.

The tendency for variance estimates to become less accurate with increasing complexity of design was referred to by Shavelson & Webb (1981) as the bandwidth-fidelity dilemma. The more factors that one includes in a study for the sake of replicating the complexity of the real world, the less confidence one can generally have in the resulting variance estimates. Due to this dilemma, experimental designs should generally be kept as simple as the research question allows, particularly with respect to crossed factors. However, the incorporation of well chosen fixed blocking variables can strengthen a design (see Samuels et al., 1991 for

examples) with little apparent adverse effects to the stability of variance estimates for the random components of the model.

Complex crossed designs not only reduce fidelity, but are also becoming less fashionable when their goal is an omnibus test of significance followed by testing for every possible effect in the hopes of breaking the .05 barrier (see Rosnow & Rosenthal, 1989). Given a simple, well defined hypothesis, such as the ones described in the preceding section, it should be possible for researchers to set up designs involving a minimum of variables. In these situations, the effort involved in addressing the fixed-effect fallacy as a threat to external validity should be minimal.

Complex designs do have their place, however. For exploratory research, the inclusion of several variables in a study can help provide a researcher with a better understanding of their relationships to one another. Furthermore, in a very large study, it may be possible to have a wide "bandwidth" and thus reflect some of the complexity in the real world, while still retaining an acceptable level of "fidelity." in the obtained variance estimates

Limitations of the Study

One of the limitations of this study is that the EMS equations from which the variance components were calculated are inappropriate to a mixed model generalizability analysis, resulting in biased variance estimates. Furthermore, the computer program utilized for the bulk of the analysis, LOGIC (Taerum, 1986), is run in single precision and reports results to only two or three decimal places, while computing in double precision and reporting

results to four decimal places is generally advocated for generalizability studies (eg., Brennan, 1983). Thus, one cannot be as confident in the numerical results as would have been the case if methods more appropriate to the mixed model had been used. However, the primary goal of this study was not the estimation of variance components or the calculation of F-ratios, but the discussion and demonstration of methods for the avoidance of the fixed-effect fallacy through an approach incorporating generalizability theory.

Another limitation is that the differences between the analysis of the unbalanced designs and artificially balanced ones, both in terms of numerical results and in terms of computing resources required, were not assessed. It is also regrettable that comparisons between computing time required for the Method III and ML analyses could not be made, because the necessary information is not reported by the VM operating system on which the ML analysis was performed.

New Algorithms for Variance Component Estimation

One of the reasons that generalizability theory is not more widely applied is that due to the requirements for relatively large sample sizes and the treatment of persons as a factor in the ANOVA, large amounts of computing resources are often required to perform generalizability analyses. As should be clear from the compromises that had to be made in the analysis of the Algebra 12 data, namely the use of a non-optimal method of moments technique and the need for reduction of the data set to perform an

ML analysis, this problem can become acute with unbalanced data. The future holds the promise of practical and attractive alternatives to method of moments approaches such as Henderson's (1953) Method III, however.

Not only are computing resources expanding and becoming more affordable, but more efficient programs for variance estimation in the unbalanced case are also being developed. In general, the most promising methods of analysis for unbalanced mixed models appear to be those adopting a maximum likelihood or Bayesian approach. As the work of Dempster, Rubin and Tsutakawa (1981) demonstrated, these approaches are not mutually exclusive. The EM algorithm developed by Dempster et al. uses a combination of Bayesian and maximum likelihood approaches to arrive at variance estimates. This approach avoids the inversion of large matrices, but is slow to converge. Longford (1987) has devised an approach by which the EM algorithm is used for the first few iterations and then the Fisher-scoring algorithm is applied, resulting in faster convergence. Advances in computing efficiency such as these will make generalizability analyses of large unbalanced more feasible in the future.

At present, the 1988 software release for BMDP contains a new program, 3V, which uses ML and REML approaches to handle the general mixed model analysis of variance (Dixon, Brown, Engelman, Hill, & Jennrich, 1988, pp. 1025-1044). This program is described as experimental as well as being "computationally intensive if the number of parameters in the model is large" (Dixon et al., 1988, p.

1026). Thus, it is not clear how practical this program will be for large scale generalizability analyses.

Implications for Future Research

Although recent advances in computational algorithms for the analysis of unbalanced mixed models are promising, there is a need to continue in the search for more efficient computing algorithms to make generalizability analyses of unbalanced data sets less costly in terms of computing resources. As new methods are developed,they will also need to be assessed, both theoretically and empirically.

Extension to More Complex Designs

In a future study planned as an extension of the one presented herein, it is hoped that one of the methods discussed in the previous section can be utilized and assessed. The logic of this extension is as follows. In many educational studies, particularly in evaluation research, complex nesting occurs on the dependent variable. For instance, students are nested within schools, which are nested within districts. Hopkins' (1984) method can be extended to the analysis of data with this structure also and this extension will be demonstrated in a subsequent study in which one of the newer methods of variance component estimation will be employed and assessed.

The large data set to be employed in this study (n = 32,000) is likely to result in a design matrix that is too large for the amount of memory available. This will necessitate the analysis of one or more subsamples of the full data set such that a manageable

design matrix is obtained. Depending on the size of manageable subsamples, replications on independent subsamples may be required because schools within districts is best considered a random facet and dispersion effects are likely to adversely affect the stability of the variance estimates for this facet.

Implications for Practice

It is hoped that researchers in education and psychology will not simply assume that the measuring device that they intend to use flawlessly reflects the construct of interest. If (a) the characteristics of the instrument used have been thoroughly described for application to the relevant populations and situations and (b) logic and previous research suggest that the fixed-effect fallacy is not a threat to external validity, then it is unlikely that the treatment of items as a random factor would add much useful information relative to the costs. If, however, a standardized measure is to be used in a non-standard situation or an untried measure is to be used, then the information gained by assessing the measurement instrument and its interactions with the other variables in the study could prove to be invaluable. Based on the information gained from this assessment, the decision can be made on the extent of the threat to external validity posed by the fixed-effect fallacy. If the relevant variance components are very small, then simplification to an essentially fixed-effects model is appropriate. The pros and cons of the generalizability approach to data analysis should be carefully

considered for the situation at hand in the decision as to whether or not it should be used.

REFERENCES

Bell, J.F. (1985). Generalizability theory: The software problem. Journal of Educational Statistics, 10, 19-29.

Bloom, B.S. (Ed.). (1956). Taxonomy of Educational Objectives, Handbook 1: The cognitive domain. New York: McGraw-Hill.

Brennan, R.L. (1983). Elements of Generalizability Theory. Iowa City: ACT Publications.

Brennan, R.L., Jarjoura, D., & Deaton, E.L. (1980). Some Issues Concerning the Estimation and Interpretation of Variance Components in Generalizability Theory. Iowa City: ACT Publications.

Brennan, R.L. & Kane, M.T. (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14, 277-289.

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E.C. Rhodes (Eds.), The marks of examiners. London: The Macmillan Company.

Cardinet, J. & Allal, L. (1983). Estimation of generalizability parameters. In L.J. Fyans, Jr. (Ed.), Generalizability Theory: Inferences and Practical Applications (pp. 17 - 48). San Francisco: Jossey-Bass

Cardinet, J., Tourneur, Y, & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13, 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. Journal of Educational Measurement, 18, 183-204.

Chastain, R.L. & Willson, V.L. (1986). Estimation of variance components using computer packages. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 12, 335-359.

Cochran, W.G. (1951). Testing a linear relation among variances. Biometrics, 7, 17-32.

Coleman, E.B. (1964). Generalizing to a language population. Psychological Reports, 14, 219-226.

Cornfield, J. & Tukey, J.W. (1956). Average values of mean squares in factorials. Annals of Mathematical Statistics, 27, 907-949.

Crocker, L & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.

Cronbach, L.J., Gleser, F.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, B. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Dempster, A.P., Rubin, D.B., & Tsutakawa, R.K. (1981). Estimation in covariance components models. Journal of the American Statistical Association, 76, 341-353.

de Finetti, B. (1964). Foresight: Its logical flaws, its subjective sources. In H. E. Kyburg & G. E. Smokler (Eds.), Studies in Subjective Probability. New York: Wiley.

Dixon, W.J., Brown, M.B., Engelman, L, Hill, M.A., & Jennrich, R.I. — (Eds.). (1988). BMDP Statistical Software Manual. Berkeley: University of California Press.

Fisher, R.A. (1925). Statistical methods for research workers. London: Oliver & Bond.

Gaylor, D.W., Lucas, H.L., and Anderson, R.L. (1970). Calculations of Expected Mean Squares by the Abbreviated Doolittle and Square Root Method. Biometrics, 26, 641-655.

Gleser, G.C., Green, B.L., & Winget, C.N. (1978). Quantifying interview data on psychic impairment of disaster survivors. The Journal of Nervous and Mental Diseases,166, 209-216.

Green, B.F. & Tukey, J. (1960). Complex analysis of variance: General problems. Psychometrika, 25, 127-152.

Hartley, H.O. (1967). Expectations, variances, and covariances of ANOVA mean squares by "synthesis". Biometrics, 23, 105-114.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72, 320-340.

Henderson, C.R. (1953). Estimation of variance and covariance components. Biometrics, 9, 226-252.

Hopkins, K.D. (1983). A strategy for analyzing ANOVA designs having one or more random factors. Educational and Psychological Measurement, 43, 107-113.

Hopkins, K.D. (1984). Generalizability theory and experimental design: Incongruity between analysis and inference. American Educational Research Journal, 21, 703-712.

Hoyt, C.J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

LaMotte, L.R. (1973). Quadratic estimation of variance components. Biometrics, 29, 311-330.

Leone, F.C. & Nelson, L.S. (1966). Sampling distributions of variance components I. Empirical studies of balanced nested designs. Technometrics, 8, 457-468.

Lindquist, E.F. (1953). Design and analysis of experiments in education and psychology. Boston: Houghton Mifflin.

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review, 72, 143-155.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika, 74, 817-827.

Malgady, R.G., Amato, J.A., & Huch, S.W. (1979). The fixed-effect fallacy in educational psychological research: A problem of generalizing to multiple populations. Educational Psychologist, 14, 70-86.

Malloy, T.E.& Kenny, D.A. (1986). The social relations model: An integrative method for personality research. Journal of Personality, 54, 199-225.

Marcoulides, G.A. (1990). An alternative method for estimating variance components in generalizability theory. Psychological Reports, 66, 379-386.

Milliken, G.A. & Johnson, D.E. (1984). Analysis of Messy Data. Belmont: Lifetime Learning Publications.

Muthen, L. (1983). The estimation of variance components for dichotomous dependent variables: Applications to test theory. Unpublished doctoral dissertation, University of California, Los Angeles.

Novick, M.R. (1975). Bayesian methods in educational testing: A third survey. In D.M.N. de Gruijter & L.J.T. van der Kamp (Eds.), Advances in Psychological and Educational Measurement (pp. 17 - 32). New York: Wiley.

Rao, C.R. (1971a). Estimation of variance and covariance components--MINQUE theory. Journal of Multivariate Analysis, 1, 257-275.

Rao, C.R. (1971b). Minimum variance quadratic unbiased estimation of variance components. Journal of Multivariate Analysis, 1, 445-456.

Rogers, W. T., & Bateson, D.J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. Applied Measurement in Education, 4(2), 159-183.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

SAS Institute Inc. (1988). SAS/STAT User's Guide, Release 6.03 Edition, Cary NC: Author.

Samuels, M.L., Casella, G., & McCabe, G.P. (1991). Interpreting blocks and random factors. Journal of the American Statistical Association, 86, 798-808.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics, 2, 110-114.

Searle, S.R. (1971). Linear Models. New York: Wiley.

Searle, S.R. (1979). Notes on variance component estimation - A detailed account of maximum likelihood and kindred methodologies. Ithica, New York: Cornell University, Biometrics Unit.

Searle, S.R. (1987). Linear Models for Unbalanced Data. New York: Wiley.

Shavelson, R.J. & Webb, N.M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics, 3, 319-346.

Smith, P.L. (1982). A confidence interval approach for variance component estimates in the context of generalizability theory. Educational and Psychological Measurement, 42, 4 459-465.

Swallow, W.H. & Monahan, J.F. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. Technometrics, 26, 47-57.

Taerum, T. (1986). Multivariate analysis of variance. Computing Services Bulletin, 20, 17-21.

Taerum, T. (1989). Efficient algorithms for analysis of variance. In K. Berk & L. Malone (Eds.), Computing Science and Statistics: 1989 Proceedings of the 21st Symposium on the Interface (pp. 475-480). Alexandria, VA: American Statistical Association.

Winer, B.J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

APPENDIX A

Table 23

Expected Mean Squares for the Balanced P:G x I Design

SOURCE	EMS
G	$\sigma^2_{ip:g} + n_p\sigma^2_{ig} + n_p\sigma^2_{p:g} + n_i n_p\sigma^2_g$
P:G	$\sigma^2_{ip:g} + n_i\sigma^2_{p:g}$
I	$\sigma^2_{ip:g} + n_p n_g\sigma^2_i$
IG	$\sigma^2_{ip:g} + n_p\sigma^2_{ig}$
IP:G	$\sigma^2_{ip:g}$

Table 24

Expected Mean Squares for the Balanced P:G x I:C Design

SOURCE	DF	EMS
G		$\sigma^2_{ip:gc} + n_p\sigma^2_{ig:c} + n_i n_c\sigma^2_{p:g} + n_i n_p n_c\sigma^2_g$
C		$\sigma^2_{ip:gc} + n_i\sigma^2_{pc:g} + n_p n_g\sigma^2_{i:c} + n_i n_p n_g\sigma^2_c$
P:G		$\sigma^2_{ip:gc} + n_i n_c\sigma^2_{p:g}$
I:C		$\sigma^2_{ip:gc} + n_p n_g\sigma^2_{i:c}$
GC		$\sigma^2_{ip:gc} + n_i\sigma^2_{pc:g} + n_p\sigma^2_{ig:c} + n_i n_p\sigma^2_{gc}$
PC:G		$\sigma^2_{ip:gc} + n_i\sigma^2_{pc:g}$
IG:C		$\sigma^2_{ip:gc} + n_p\sigma^2_{ig:c}$
IP:GC		$\sigma^2_{ip:gc}$

APPENDIX A (CONTINUED)

Table 25

Expected Mean Squares for the Balanced P:G x I:L Design

SOURCE	EMS
G	$\sigma^2_{ip:gl} + n_p\sigma^2_{ig:l} + n_in_l\sigma^2_{p:g} + n_in_pn_l\sigma^2_g$
L	$\sigma^2_{ip:gl} + n_i\sigma^2_{pl:g} + n_pn_g\sigma^2_{i:l} + n_in_pn_g\sigma^2_l$
P:G	$\sigma^2_{ip:gl} + n_in_l\sigma^2_{p:g}$
I:L	$\sigma^2_{ip:gl} + n_pn_g\sigma^2_{i:l}$
GL	$\sigma^2_{ip:gl} + n_i\sigma^2_{pl:g} + n_p\sigma^2_{ig:l} + n_in_p\sigma^2_{gl}$
PL:G	$\sigma^2_{ip:gl} + n_i\sigma^2_{pl:g}$
IG:L	$\sigma^2_{ip:gl} + n_p\sigma^2_{ig:l}$
IP:GL	$\sigma^2_{ip:gl}$

APPENDIX A (CONTINUED)

Table 26

Expected Mean Squares for the Balanced P:G x I:CL Design

SOURCE	EMS
G	$\sigma^2_{ip:gcl} + n_p\sigma^2_{ig:cl} + n_i n_l n_c \sigma^2_{p:g} + n_i n_p n_c n_l \sigma^2_g$
L	$\sigma^2_{ip:gcl} + n_i n_c \sigma^2_{pl:g} + n_p n_g \sigma^2_{i:cl} + n_i n_p n_g n_c \sigma^2_l$
C	$\sigma^2_{ip:gcl} + n_i n_l \sigma^2_{pc:g} + n_p n_g \sigma^2_{i:cl} + n_i n_p n_g n_l \sigma^2_c$
P:G	$\sigma^2_{ip:gcl} + n_i n_c n_l \sigma^2_{p:g}$
I:LC	$\sigma^2_{ip:gcl} + n_p n_g \sigma^2_{i:lc}$
GL	$\sigma^2_{ip:gcl} + n_p\sigma^2_{ig:cl} + n_i n_c \sigma^2_{pl:g} + n_i n_c n_p \sigma^2_{gl}$
GC	$\sigma^2_{ip:gcl} + n_p\sigma^2_{ig:cl} + n_i n_l \sigma^2_{pc:g} + n_i n_l n_p \sigma^2_{gc}$
CL	$\sigma^2_{ip:gcl} + n_i\sigma^2_{pcl:g} + n_p n_g \sigma^2_{i:cl} + n_i n_p n_g \sigma^2_{cl}$
GCL	$\sigma^2_{ip:gcl} + n_i\sigma^2_{pcl:g} + n_p\sigma^2_{ig:cl} + n_i n_p \sigma^2_{gcl}$
PL:G	$\sigma^2_{ip:gcl} + n_i n_c \sigma^2_{pl:g}$
PC:G	$\sigma^2_{ip:gcl} + n_i n_l \sigma^2_{pc:g}$
IG:LC	$\sigma^2_{ip:gcl} + n_p\sigma^2_{ig:cl}$
CLP:G	$\sigma^2_{ip:gcl} + n_i\sigma^2_{clp:g}$
IP:GCL	$\sigma^2_{ip:gl}$

APPENDIX B

Calculation of Biased and Unbiased Estimates of $\sigma^2_{p:g}$ in the P:G x I:C Design

Calculation of the Biased Estimate for $\sigma^2_{p:g}$

$E(MS_{P:G}) = \sigma^2_{ip:gc} + 12.491\sigma^2_{pc:g} + 55.000\sigma^2_{p:g}$

$55.000\sigma^2_{p:g} = E(MS_{P:G}) - E(MS_{IP:G})$

$- 12.491/10.627[E(MS_{PC:G}) - E(MS_{IP:G})]$

$55.000\sigma^2_{p:g} = 0.91 - 0.18 - 1.175(0.21 - 0.18)$

$\sigma^2_{p:g} = 0.694/55.000 = 0.01262$

Calculation of the Unbiased Estimate for $\sigma^2_{p:g}$

$E(MS_{P:G}) = \sigma^2_{ip:gc} + 55.000\sigma^2_{p:g}$

$55.000\sigma^2_{p:g} = E(MS_{P:G}) - E(MS_{IP:G})$

$\sigma^2_{p:g} = (0.91 - 0.18)/55.000 = 0.01327$

APPENDIX C

Satterthwaite (1946) devised a procedure by which an approximate confidence interval for variance components can be computed. The approximate 100(P)% confidence interval is:

$$\left[\frac{\hat{\sigma}^2(\alpha|M)\nu}{\chi^2_U{}^{\nu}} \leq \sigma^2(\alpha) \leq \frac{\hat{\sigma}^2(\alpha|M)\nu}{\chi^2_L{}^{\nu}}\right]$$

where $\hat{\sigma}^2\alpha|M$ is the estimated variance given the model M, j indexes the mean squares that enter into the estimation of $\hat{\sigma}^2\alpha|M$, and f_j is the coefficient of MS_j in the linear combination of MS_j that gives $\hat{\sigma}^2\alpha|M$ (see Table 2 and Appendix B). The terms $\chi^2_U{}^{\nu}$ and $\chi^2_L{}^{\nu}$ are the upper, U = (1 + P)/2, and lower, L = (1 - P)/2, percentage points of the χ^2 distribution with

$$\nu = \frac{(\sum_j f_j MS_j)^2}{\sum_j (f_j MS_j)^2/df_j}$$

"effective" degrees of freedom. For a more extensive discussion of this procedure, see Brennan (1983, pp. 101 - 104, 137).